AD-A058 082 NATIONAL AVIATION FACILITIES EXPERIMENTAL CENTER ATL--ETC F/G 5/9
DEVELOPMENT OF A PERFORMANCE CRITERION FOR AIR TRAFFIC CONTROL --ETC(U)
JUL 78 E P BUCKLEY, K HOUSE, R ROOD

UNCLASSIFIED FAA-NA-78-9 FAA-RD-78-71 NL

1 OF 2
ADA
058082

Report No. FAA-RD-78-71

DEVELOPMENT OF A PERFORMANCE CRITERION FOR AIR TRAFFIC CONTROL PERSONNEL RESEARCH THROUGH AIR TRAFFIC CONTROL SIMULATION

Edward P. Buckley
Kenneth House
Richard Rood

JULY 1978

FINAL REPORT

Document is available to the U.S. public through the National Technical Information Service, Springfield, Virginia 22161.

Prepared for

U.S. DEPARTMENT OF TRANSPORTATION
FEDERAL AVIATION ADMINISTRATION
Systems Research & Development Service
Washington, D.C. 20590

NOTICE

The United States Government does not endorse products or manufacturers. Trade or manufacturer's names appear herein solely because they are considered essential to the object of this report.

Technical Report Documentation Page

1. Report No.
(18) FAA-RD-78-71 (19)

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
(6) DEVELOPMENT OF A PERFORMANCE CRITERION FOR AIR TRAFFIC CONTROL PERSONNEL RESEARCH THROUGH AIR TRAFFIC CONTROL SIMULATION.

5. Report Date
(11) Jul[redacted] [redacted]78 (12) 99p

6. Performing Organization Code

7. Author(s)
(10) Edward P. Buckley,
Kenneth House ~~and~~ Richard Rood

8. Performing Organization Report No.
(14) FAA-NA-78-9

9. Performing Organization Name and Address
Federal Aviation Administration
National Aviation Facilities Experimental Center
Atlantic City, New Jersey 08405

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
216-101-100

12. Sponsoring Agency Name and Address
U.S. Department of Transportation
Federal Aviation Administration
Systems Research and Development Service
Washington, D.C. 20590

13. Type of Report and Period Covered
(9) Final rept.
Jan[redacted] [redacted]75 - Oct[redacted] [redacted]77

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This report gives the theory of an approach to objective measurement of the radar control performance of air traffic controllers, by means of air traffic control simulation exercises. A set of objective measurements developed for the NAFEC Air Traffic Control Simulation Facility is described. The relevance of this same measurement technique for either evaluating new systems (when the same or similar controller teams are functioning) or for evaluating various controller individuals or teams (when they are using the same system to control traffic) is discussed. Other applications are also described. The ability of the simulator to repeatedly present the same traffic samples is stressed as a means of accumulating comparable and normative data. The need for basic experimentation for validation of the test measurement system and to develop further knowledge and understanding of the measurements is recognized. A relatively small keystone experimental design is described and recommended as the first essential step for all possible applications. The availability of adequate numbers of controllers as subjects is recognized as the major problem to be overcome. Development of a means for transmitting tests originating in NAFEC's simulator to field sites is recommended.

17. Key Words
Personnel Research
Air Traffic Control
Controller Performance Measurement
Performance Criteria
Air Traffic Control Simulation

18. Distribution Statement
Document is available to the public through the National Technical Information Service, Springfield, Virginia 22151

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	94	

Reproduction of completed page authorized

240 550

METRIC CONVERSION FACTORS

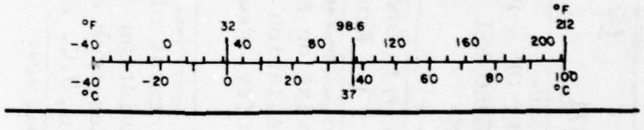
Approximate Conversions to Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
		AREA		
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
		MASS (weight)		
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
		VOLUME		
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
		TEMPERATURE (exact)		
°F	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C

*1 in = 2.54 (exactly). For other exact conversions and more detailed tables, see NBS Misc. Publ. 286, *Units of Weights and Measures*, Price $2.25, SD Catalog No. C13.10:286.

Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
		AREA		
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
		MASS (weight)		
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
		VOLUME		
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
		TEMPERATURE (exact)		
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	°F



PREFACE

There is an identified crucial need in the FAA for an objective method of measuring air traffic controller performance for many purposes, including improvement of controller selection and training. The work on controller performance measurement carried out at NAFEC over the last few years has been aimed at providing such a capability. In addition, the test methodology under development is directly usable for other applications including: the evaluation of equipment, software, and new air traffic systems; verification of mathematical models used to predict future systems effects, etc. This is a final report, in that NAFEC's controller performance measurement work has reached a point beyond which further progress is not possible without appropriate, but unavailable subjects; and therefore the project was terminated. The effort, however, remains incomplete, in the sense that the planned experimentation to validate and refine the concept was not accomplished. It is clear, to the writers of the report at least, that the work will be taken up again at some point because of the FAA needs it can satisfy. This report, therefore, is an attempt to clearly explain the basic concept and provide a record of what was accomplished and what directions seemed worthwhile, so that when an adequate means of obtaining the necessary subjects is established, it will be easier to begin again.

The authors recognize and acknowledge the invaluable assistance and support provided by a large number of individuals. Special thanks, however, must go to: Mr. Richard Algeo, who held primary responsibility for accomplishing the necessary software development; Mr. Bernard Goldberg, who provided much needed air traffic controller input in the beginning stages of the effort; Mr. Thomas Morgan of Computer Sciences Corporation, who prepared the experimental design; Dr. Albert E. Beaton of the Department of Statistics at Princeton University, who served as a skillful and friendly consultant on the experimental design; Mr. Raymond H. Ratzlaff, Chief, Human Engineering Branch, NAFEC, without whose encouragement we wouldn't have gotten far; and Mr. Ben Wenning of FAA Systems Research and Development Service, for cheerful encouragement during some dark hours.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF ILLUSTRATIONS

INTRODUCTION

OBJECTIVE.

The objective of the effort described herein was to develop an objectively scored performance measurement system for radar control performance of air traffic controllers.

BACKGROUND.

In January 1975, a study of air traffic control specialist (ATCS) training, performed by the Institute for Defense Analyses (IDA), presented a strong case for several research projects that would support improvements in selection, performance measurement, and training of air traffic controllers (reference 2). Included in the study was a recognition of the need for ATCS performance measures and recording and for analytical capability to provide training scores. Accomplishment was identified as being dependent upon work underway at the National Aviation Facilities Experimental Center (NAFEC). The House Committee on Government Operations approved and adopted a report in January 1976 (reference 3) which referenced the IDA report and emphasized the need for support to the work being done in this area by NAFEC. Appendix A contains pertinent excerpts from these documents.

As a result of the IDA study, a Federal Aviation Administration (FAA) Engineering and Development (E&D) Working Group on ATCS training was established, composed of representatives from the Office of Systems Engineering Management (OSEM), the Systems Research and Development Service (SRDS), NAFEC, the Air Traffic Service (ATS), the Office of Aviation Medicine (OAM), and the Office of Personnel and Training (OPT). This group had a twofold objective:

1. To determine what role, if any, the FAA E&D organization could play in improving controller selection, training, and evaluation, and

2. To plan the program to fulfill that role.

The working group effort resulted in a request for support in the development of ATCS training from the Associate Administrator for Administration, AAD-1, to the Associate Administrator for Engineering and Development, AED-1 (appendix B), and an E&D Program Plan for ATCS Personnel Support (reference 1).

TEST DEVELOPMENT HISTORY.

As indicated in the IDA report (reference 2) a sound, but minimal, effort in the area of performance measures was already underway at NAFEC as a result of previous work and requests. In 1969, a project on the effects of aging on air traffic controllers, carried out at NAFEC, concluded that there were significant and measurable differences in capability among journeymen air traffic controllers (reference 4). The Air Traffic Controller Career Committee also produced a report (The Corson Report) in January 1970 which recommended the use of the NAFEC dynamic simulator to develop systematic and objective means for evaluation of the proficiency of controllers on the job (reference 5). In 1972, a letter from the Office of Manpower (reference 6) requested SRDS: "To establish a resource center for continuing development of methodology for objective measurement of controller performance using dynamic simulation. The instruments of measurement and data developed will provide a valuable resource available for personnel research, determination of training methods, training evaluation, and may also be helpful in the evaluation of individual controller proficiency. ...By use of experimental testing, a set of norms should also be developed, including information on the elements of the traffic situation which generated in various degrees the relative difficulty for the controller of different traffic sets and situations." The result was continuation of the research at NAFEC.

The new program, established in 1976, gave higher priority and increased stature to the work effort and required "... field evaluation involving a large sample of journeyman controllers" (reference 1). It also addressed the perplexing problem of obtaining the required large numbers of appropriate subjects, by stipulating the fabrication of a mobile test laboratory. This would, in effect, bring the laboratory to the subjects by receiving and displaying data remoted from the Air Traffic Control Simulation Facility (ATCSF), previously known as the Digital Simulation Facility (DSF), at NAFEC. A plan for developing such a capability was prepared by NAFEC. However, in evaluating both this plan and its alternative, which was to transport large numbers of controllers to NAFEC, the FAA decided to defer all experimentation until the new air traffic control (ATC) simulator (to be known as the Radar Training Facility (RTF)) is installed at the FAA Academy in Oklahoma City, Oklahoma, and becomes available for this type of testing, in approximately 1981. No technical objection to the NAFEC approach nor to the work accomplished was found; however, the result of these deliberations was the termination of the efforts associated with this project.

This report describes the rationale, laboratory, tests, measurements and experimental design established at NAFEC for this project up to the point of termination. The possible variations and applications of such a test capability are also discussed.

TEST RATIONALE

A detailed discussion of the rationale followed in developing the controller performance measurement concept covered in this report is given in the paper "Systems Performance Measurements and Individual Performance Measurements" (appendix C) presented to the Second International Learning Technology in Orlando, Flordia, on February 14, 1978. In order to insure a complete understanding of the effort described in the following sections, it is recommended that appendix C be read prior to proceeding. However, a brief digest of the test rationale follows here.

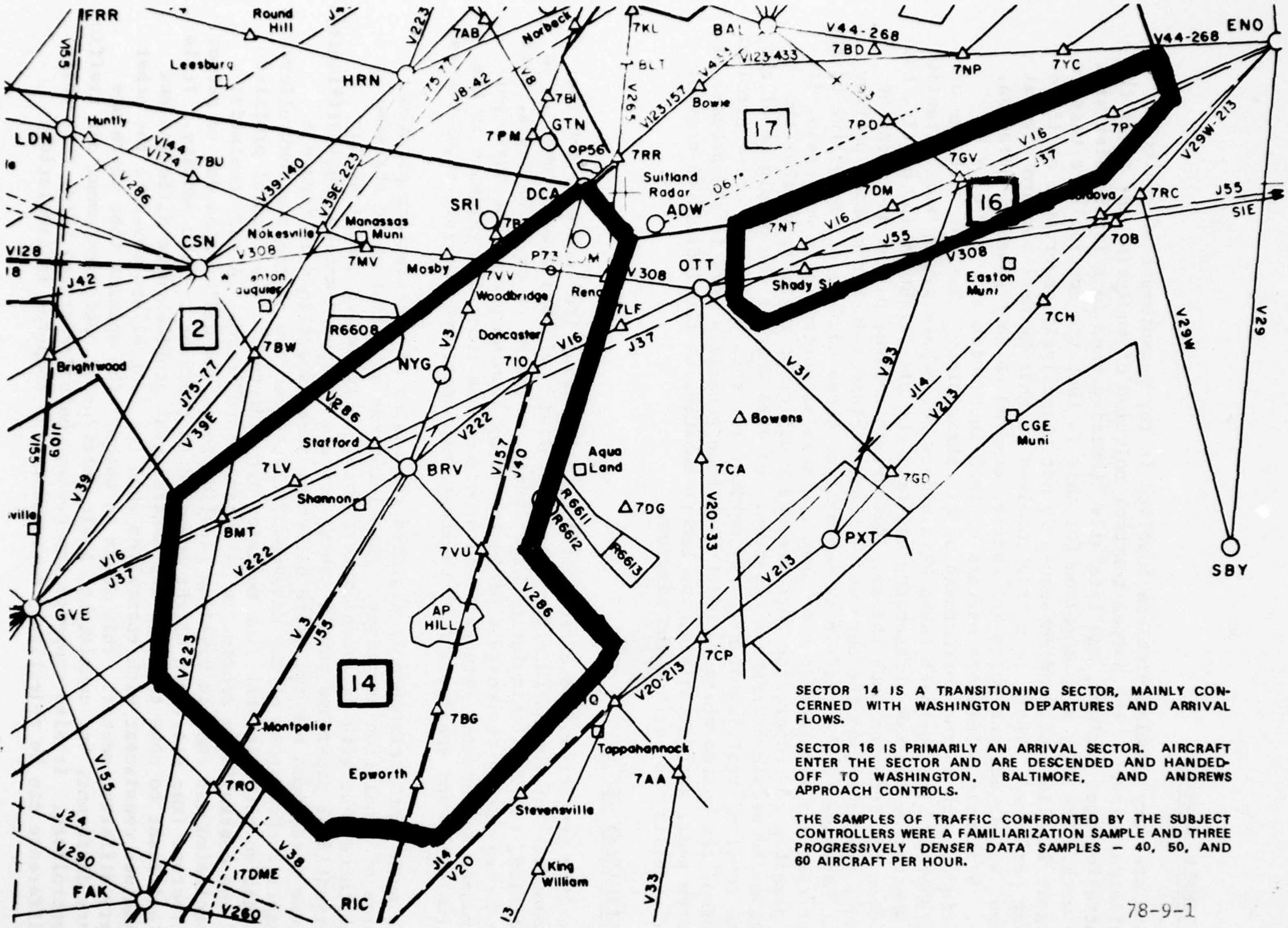
RATIONALE SUMMARY.

There are very complex problems involved in the measurement of controller proficiency level. The normal textbook solution of analyzing the tasks the controller must perform, and carefully itemizing them in a little list will not work. The very simple reason for this is that the controller's task is largely intellectual, and consequently not objectively observable as it is being performed. There is a trap involved in that he does perform several manual verbal actions which are easily measured as to frequency, duration, etc. Similarly, there are matters of technique which are very easy to critique, such as the correctness of his phraseology. But these things are peripheral. The approach taken herein does not negate giving some attention to whether he can use phraseology and fill out his flight strips neatly, but it does not focus on such things. Rather, it focuses on the decisions he makes, or rather, on the outcome of these decisions in terms which portray what happened to the traffic as a result of those decisions. By virtue of the fact that the simulator can present the same situation again and again, it is possible to compare the outcomes under different regimes, whether the regimes be people or teams or systems, and also whether the people be of the same level of training or not. By making, for this purpose, a one-controller system, the system measures which have been generally used in the process of system test, such as delays, can become measures of the individual controller's ability to perform the mission.

PRELIMINARY TESTS.

The 1969 project on the effects of controller aging (reference 6) set the groundwork for the NAFEC efforts on controller performance measurement. In addition, two small simulation experiments, referred to as PROBE tests, were conducted during the current project. These also supported the basic performance measurement rationale and provided experience and guidance in the formulation and application of the test, measures, and analysis requirements.

One such test (reference 7) explored the basics of constructing parallel forms of simulation exercises. Parallel forms of a test cover the same substance, but with different material (e.g., items and questions in conventional paper and pencil tests) and are of approximately equal difficulty. These tests used the two quite different sectors outlined in figure 1. Figure 2 shows some of the measures collected on two of the six controllers tested on both sectors. The measures are depicted in the form of profiles of the controller's scores, in "standard score" terms. A very tentative conclusion was that in building parallel forms the differences due to sector structure (map etc.) are slight compared to those of traffic density. It is significant to note that the two controllers' profiles were different and remained consistent with themselves when the controllers moved to the other very different sector. This seems to support the opinion that there are very different, but consistent, basic patterns of operation among air traffic controllers. It also supports the concept being described here, that this difference can be objectively measured by the computer in simulation.

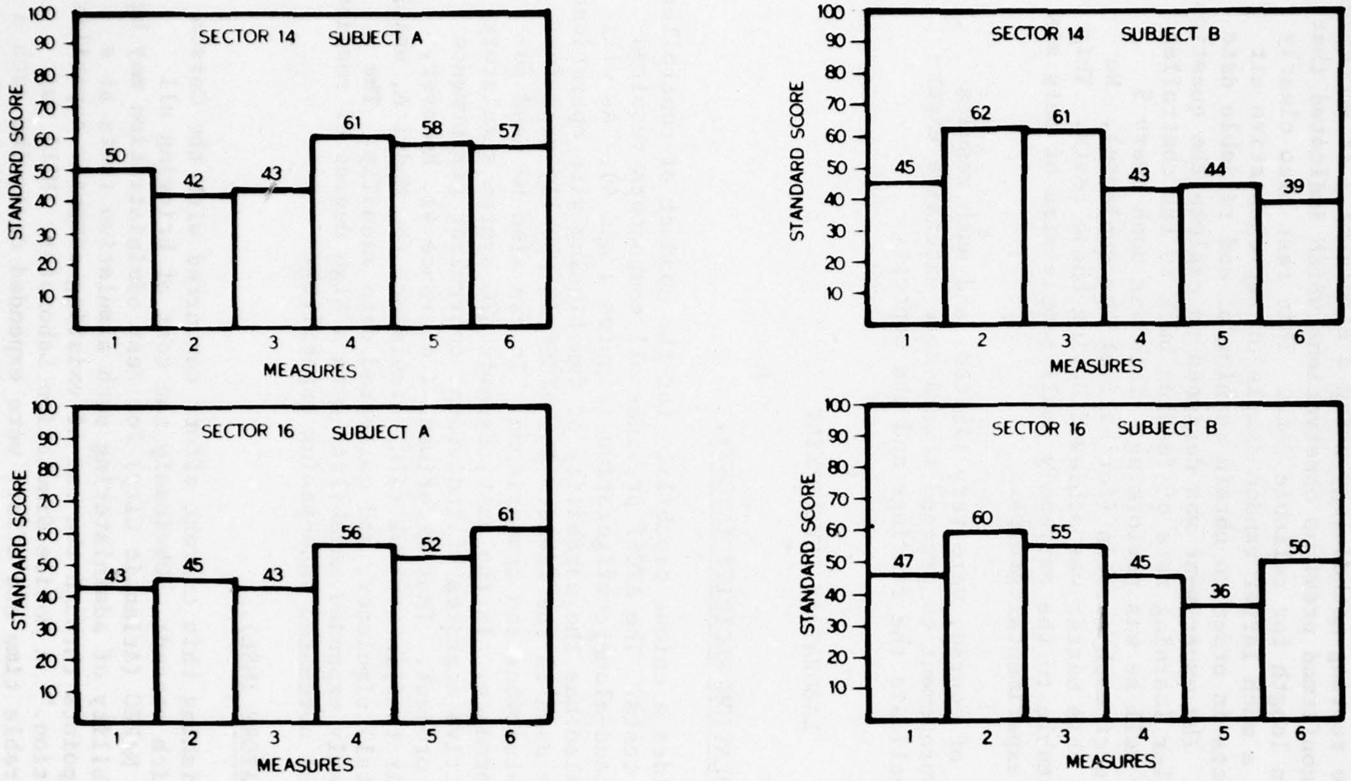


SECTOR 14 IS A TRANSITIONING SECTOR, MAINLY CONCERNED WITH WASHINGTON DEPARTURES AND ARRIVAL FLOWS.

SECTOR 16 IS PRIMARILY AN ARRIVAL SECTOR. AIRCRAFT ENTER THE SECTOR AND ARE DESCENDED AND HANDED-OFF TO WASHINGTON, BALTIMORE, AND ANDREWS APPROACH CONTROLS.

THE SAMPLES OF TRAFFIC CONFRONTED BY THE SUBJECT CONTROLLERS WERE A FAMILIARIZATION SAMPLE AND THREE PROGRESSIVELY DENSER DATA SAMPLES — 40, 50, AND 60 AIRCRAFT PER HOUR.

FIGURE 1. SECTOR MAPS USED FOR EARLY TESTS



MEASURES:

1 - NUMBER OF CONFLICTIONS/NUMBER OF AIRCRAFT HANDLED
2 - NUMBER OF DELAYS/NUMBER OF AIRCRAFT IN SAMPLE
3 - CUMULATIVE DELAY TIME/NUMBER OF AIRCRAFT IN SAMPLE
4 - NUMBER OF COMPLETED FLIGHTS/NUMBER OF COMPLETABLE FLIGHTS
5 - NUMBER OF AIRCRAFT HANDLED/NUMBER OF AIRCRAFT IN SAMPLE
6 - CORRELATION HANDLED/DELAY TRANSFORM

78-9-2

FIGURE 2. STANDARD SCORE PROFILES OF TWO CONTROLLERS ON TWO SECTORS

Another test (reference 8) was conducted to aid the process of developing and refining an objective testing methodology using a simulator. It succeeded in doing this. It also confirmed previous observations which indicated that 2 hours was a minimum run length for reliable data. This test also clearly established the need for a much larger random sample of representative air traffic controller subjects in order to obtain meaningful and reliable data from future experiments. The experiment was designed to explore the question of the effect on controller learning rate of feeding back to the controller quantitative data on how well he was performing. This was done every 5 minutes during the course of each session (for half of the subjects). No definitive conclusion on this matter was achieved during these tests. This is most probably attributable to the extremely small sample size of only six subjects in each of two experimental groups.

These preliminary tests, of course, were very limited, and such results served primarily as encouragement to proceed toward more extensive experimentation to properly validate the findings and the concept.

LABORATORY DEVELOPMENT

AIR TRAFFIC CONTROL SIMULATION FACILITY (ATCSF).

The ATCSF at NAFEC provides a unique capability for the conduct of controller performance measurement tests. The ATCSF provides all components required for such testing in a stand-alone configuration (figures 3 and 4). As will be discussed later, it also has the capability of functioning with operational components such as those used in the field. The ATCSF is usually used for equipment and system evaluations and comparisons. It can also be used to measure controller performance. In the past, dynamic ATC system simulators have been used for objective measures of individual controller performance only once prior to this project. That experiment (reference 4), however, used NAFEC's previous air traffic control (ATC) simulator, the Model A, which was an analog (not digital) simulator, and collected data manually. The ATCSF provides considerably expanded capability with a high degree of realism (reference 9) and digital automated data-taking capability.

MOBILE SIMULATION LABORATORY (MSL).

The program plan establishing this current effort concurred with the Corson Report (reference 5) which stated: "Obviously the cost of bringing all controller personnel to NAFEC (Atlantic City) for test administration may be prohibitive. The possibility of administering such simulation tests at a number of geographical points through the use of existing computer capability deserves prompt exploration." A Mobile Simulation Laboratory (MSL) was called for, and considerable time and effort were expended designing such a laboratory. Plans were made for the acquisition of a large truck-trailer to carry the controller positions and a processor. Operational Plan View Displays (PVD) used in en route facilities were found to be very expensive

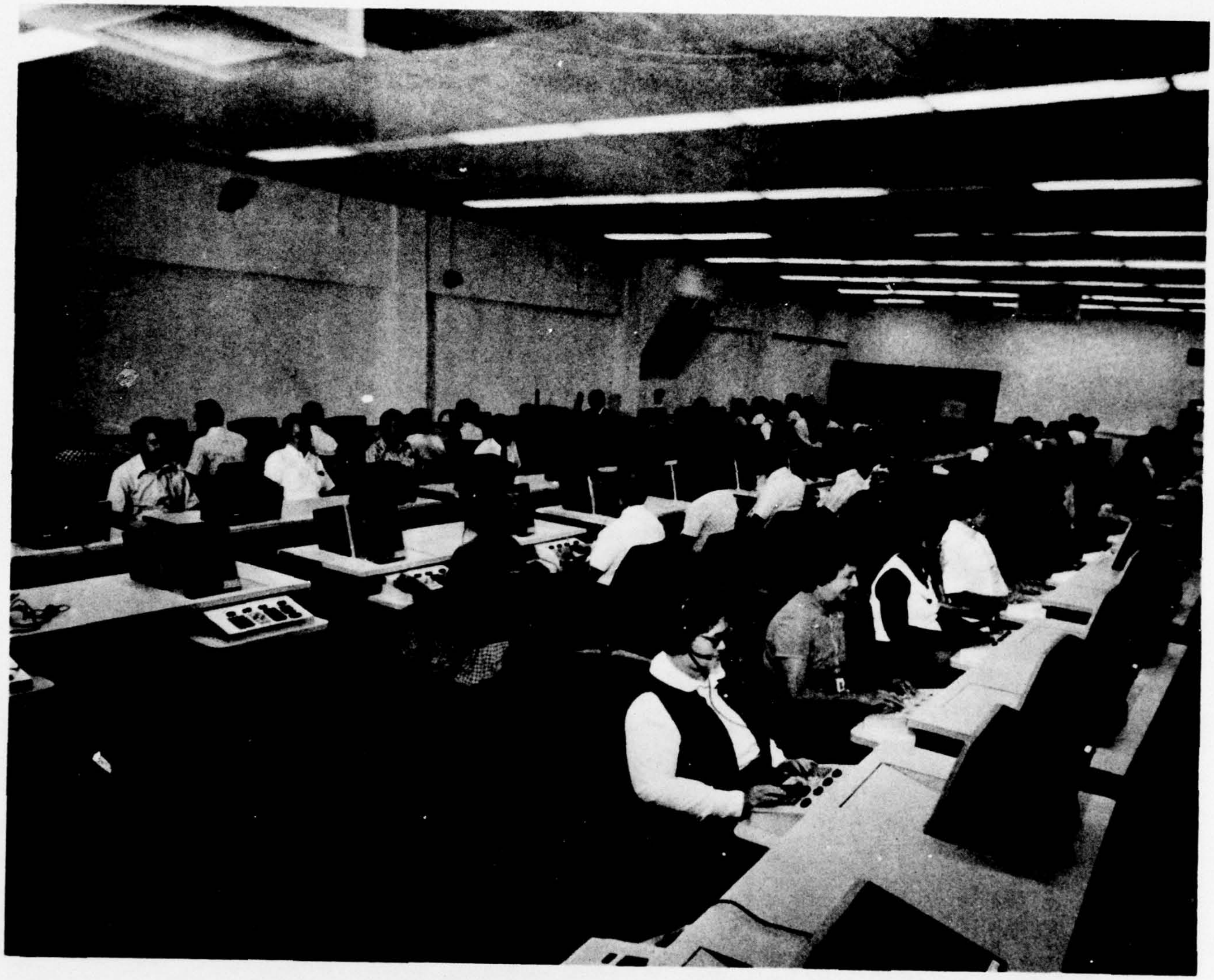


FIGURE 3. ATCSF SIMULATOR PILOTS

78-9-3

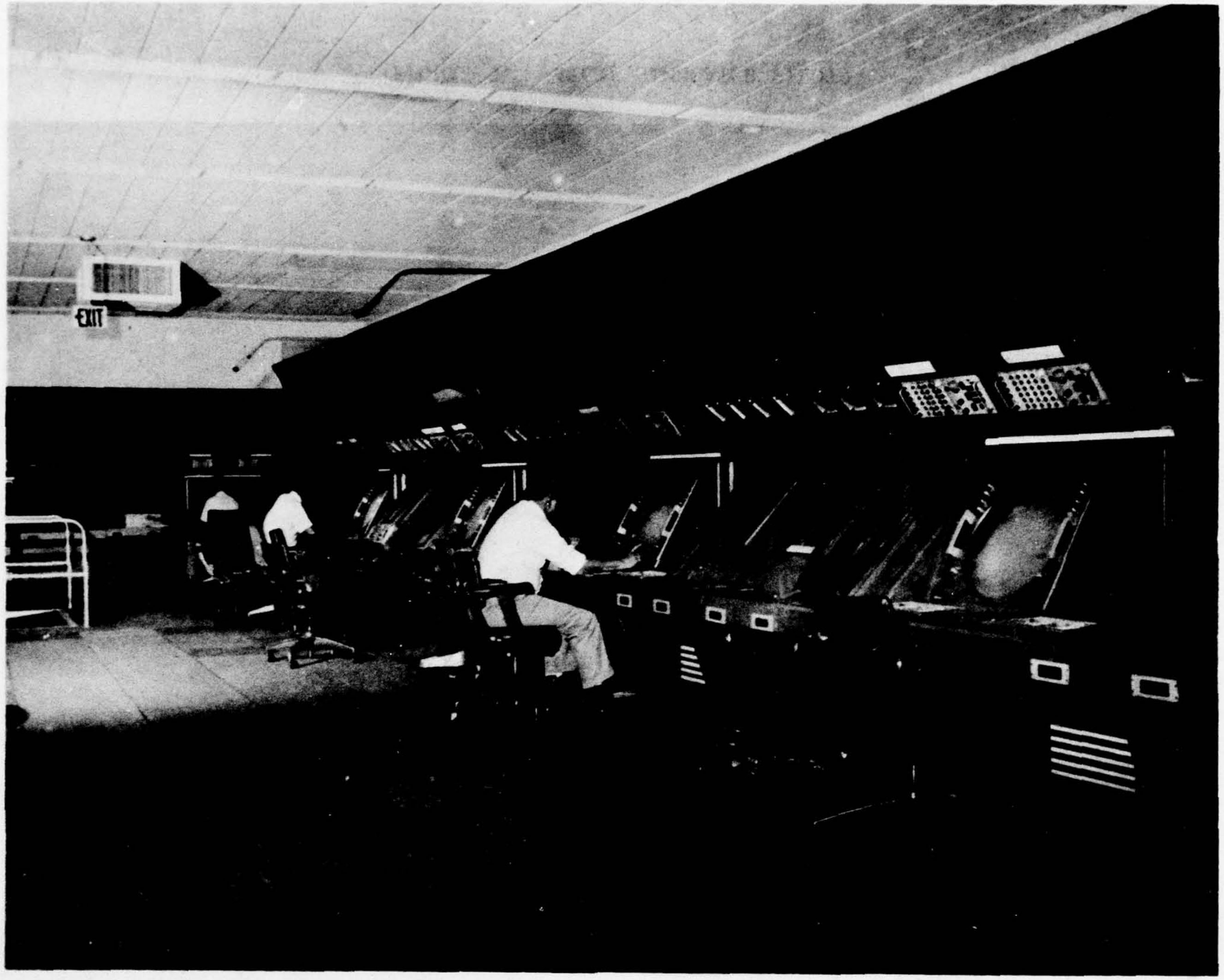


FIGURE 4. ATCSF CONTROLLER LABORATORY

78-9-4

and impractical for such a mobile and limited-size environment; therefore, other, cheaper, less sensitive displays, usable for either terminal or en route type tests, were found. Since the MSL could contain four such controller positions, a modification was made to the ATCSF software to permit four identical tests to be administered concurrently and independently. Each position could perform as a completely independent sector. Simulated interaction with adjacent facilities was provided by ghost controller positions in the ATCSF, interfaced with the subject controllers via interphone. Transmission of the simulated radar signals, generated by the ATCSF, and the voice messages to and from simulator pilots, subject controllers, and support positions would have been via telephone long-lines or by satellite.

The National Aviation and Space Administration (NASA), Goddard Space Flight Center, was interested in this effort as a possible means of providing them with additional experience in special applications of satellite communications. Increasing costs of the long-line communications made their offer of cooperation and assistance especially attractive. Planning and scheduling problems also became a major consideration in dealings with the long-line communications companies, which severely restricted the movement of a laboratory of this type. Installation of the long-line terminals at the various facilities requires a long lead time. The satellite approach offered the possibility of increased flexibility (mobility) and reliability.

An MSL of the type described has the potential of greatly enhancing the test and evaluation capability of the FAA through the addition of such a versatile tool to the NAFEC inventory. There is a recognized need for the inclusion of current air traffic controllers in the development, test, and evaluation process for new equipment, software, or systems (reference 10). An adequate method of maintaining such a group at NAFEC has not yet been achieved. Bringing in an adequate number of controllers from the field is very costly and disruptive of operations in the field.

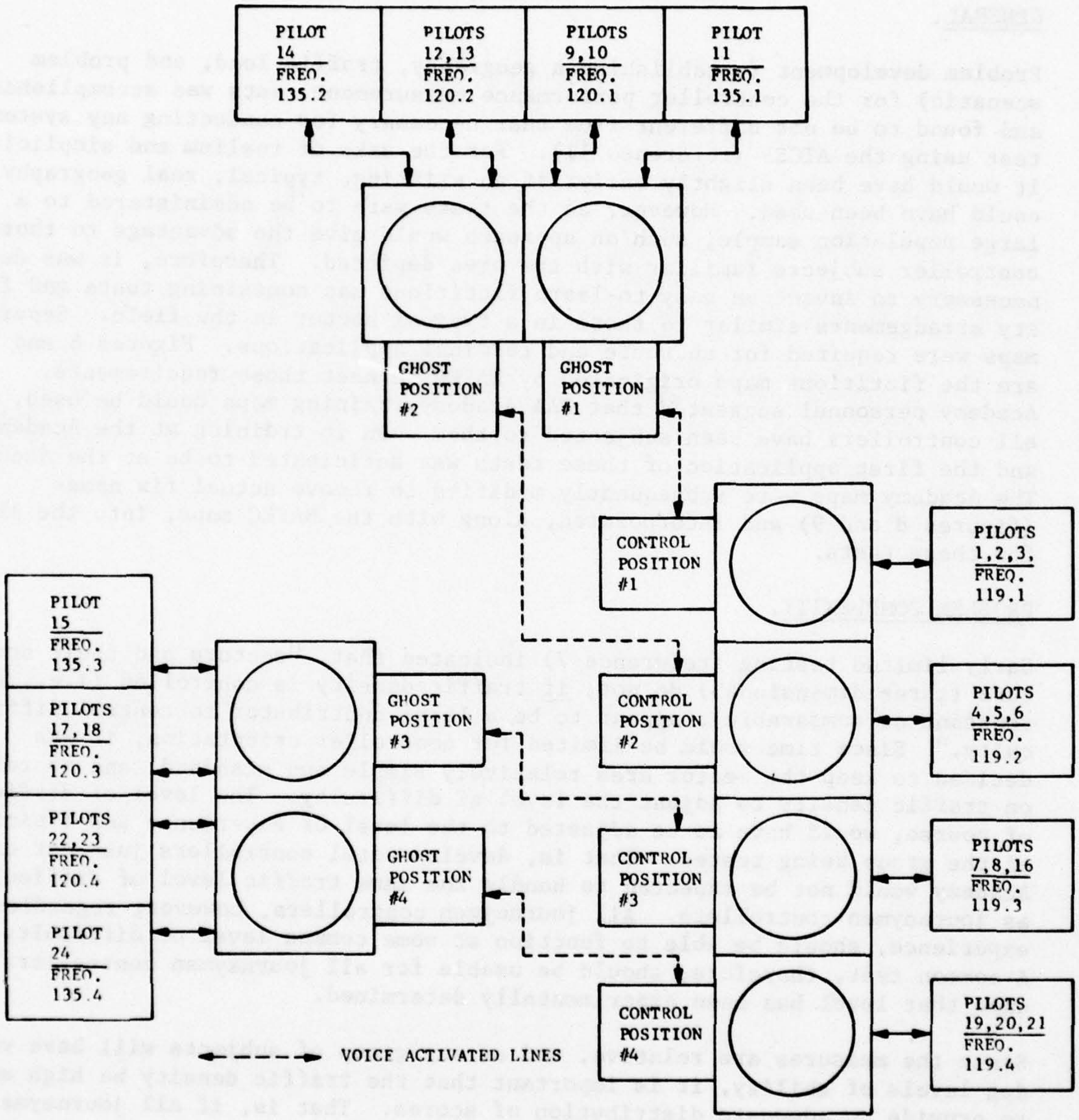
The MSL would provide a means of reasonably bringing the test to any desired number of appropriate subjects with virtually no impact on field operations. Although long-line communications could be utilized when reasonably priced satellite usage is not available, future use of satellites as the FAA communications media can easily be envisioned. The tests planned here would have provided, as a secondary result, proof of the concept. Expert authorities from NASA, the FAA Aerosat Program, NAFEC, and Westinghouse have confirmed that the plans being formulated for use of the satellite with the MSL were not only within the state-of-the-art, but were similar to other systems in regular routine operation, and posed no special technical risk.

A joint NAFEC/NASA effort was, therefore, planned. NASA was to provide the use of the ATS-6 satellite for at least 1 year, pending discussion of satellite schedule details. The loan of a truck-trailer and ground equipment from NASA was also possible. Additional surplus satellite communications ground equipment was available from the Transportation System Center, if desired. Four displays for the individual four positions and some computer equipment for interface processing in the MSL were also purchased. Software, previously developed in the ATCSF for similar kinds of external interface, required relatively minor modifications for use with this project. Software development for the interface processor was started. The decision not to proceed with the MSL stopped the MSL development and interface work. No other significant changes to the basic system design, however, were required, as all test development was already using four controller positions in the ATCSF.

SIMULATION STANDARDIZATION AND REALISM.

One of the most important advantages of using a simulator for the conduct of ATC tests is the ability to maintain and repeat absolutely standard conditions at all times to all subjects. Obviously once the test is underway, the aircraft traffic would be dynamically responsive to the control decisions made by the subject controller; however, all aspects of the environment and traffic introduced into the problem would be identical at all positions. The actions of the ghost positions (simulated adjacent sectors or facilities) would be tightly scripted to insure consistent (though operationally reasonable) actions and reactions to subject controller requests, in all cases. In order to insure that only the subject controller's abilities would be measured, the subject controller would be required to perform all sector functions alone, without the aid of an assistant controller, tracker, or coordinator.

The obvious deviations from the real world control environment cannot be taken lightly. For such tests, realism is of considerable importance. However, experience has shown that realism can be compromised to some extent without jeopordizing the results. The IDA report (reference 1) found that: "Only that part of air traffic control that is a precise sensory-motor skill requires high fidelity in the simulation. On the other hand, if the critical skills are mostly in the areas of decision making and communication, completeness, rather than precise realism of the display on the scope, will probably be most significant. In the final analysis, the validity of a simulation has to be proven by research and experiment." In this case, a controller's independent ability to properly control air traffic would be measured. The focal point of such ability is the radar position; therefore, the radar display was designed to be virtually identical to that presented on an en route PVD or terminal Data Entry and Display System (DEDS). The entry keyboard was not identical to an operational system, but skill in keyboard entries was not the concern of such tests and such entries were limited almost entirely to simple INITIATE HANDOFF, ACCEPT HANDOFF, MOVE DATA BLOCK, etc., actions. Flight plan entries, amendments, etc., were not required. Equipment for communications with adjacent facilities, sectors, pilots, etc., were not identical, but were very similar to real life actions and controls. The flight strips required were to be printed prior to the start of the tests. The controller laboratory configuration is shown in figure 5. Considering that the goal was to provide a controlled, standardized means of measuring a single subject (developmental or journeyman) controller's radar control decision-making ability, relative to others in his peer group, a high degree of realism was achieved. Where deviations were necessary, they were relatively simple to adjust to, were consistent with all subject controllers, and did not detract from, nor distort, the essential control task.



78-9-5

FIGURE 5. ATCSF CONTROLLER LABORATORY CONFIGURATION

PROBLEM DEVELOPMENT

GENERAL.

Problem development (establishing a geography, traffic load, and problem scenario) for the controller performance measurement tests was accomplished and found to be not different from that necessary for conducting any system test using the ATCSF (reference 11). For the sake of realism and simplicity, it would have been slightly easier if an existing, typical, real geography could have been used. However, as the tests were to be administered to a large population sample, such an approach would give the advantage to those controller subjects familiar with the area depicted. Therefore, it was deemed necessary to invent an easy-to-learn fictitious map containing route and facility arrangements similar to those in a typical sector in the field. Separate maps were required for en route and terminal applications. Figures 6 and 7 are the fictitious maps originated by NAFEC to meet those requirements. Academy personnel suggested that FAA Academy training maps could be used, as all controllers have been subjected to them when in training at the Academy, and the first application of these tests was anticipated to be at the Academy. The Academy maps were subsequently modified to remove actual fix names (figures 8 and 9) and incorporated, along with the NAFEC maps, into the ATCSF for these tests.

PROBLEM COMPLEXITY.

Early limited testing (reference 7) indicated that, "sectors and their structure (three dimensional) do not, if traffic density is controlled (i.e., kept constant or comparable), appear to be a large contributor to control difficulty." Since time would be limited for controller orientation, it was decided to keep the sector area relatively simple and standard, and to rely on traffic density to adjust the level of difficulty. The level of difficulty, of course, would have to be adjusted to the level of experience and training of the group being tested. That is, developmental controllers just out of the Academy would not be expected to handle the same traffic level of difficulty as journeymen controllers. All journeymen controllers, however, regardless of experience, should be able to function at some common level of difficulty. A common test, therefore, should be usable for all journeymen controllers once that level has been experimentally determined.

Since the measures are relative, and as any group of subjects will have varying levels of ability, it is important that the traffic density be high enough to provide an adequate distribution of scores. That is, if all journeymen controllers could handle a certain traffic sample with absolutely no conflictions or aircraft delays, then all of the journeymen controllers would receive the same score, and no information would be yielded. Similarly, if every journeyman received the same huge scores, that traffic sample would be pointless. The level chosen for journeymen, therefore, should perhaps be higher than the normal realistic peak level traffic for one controller, but

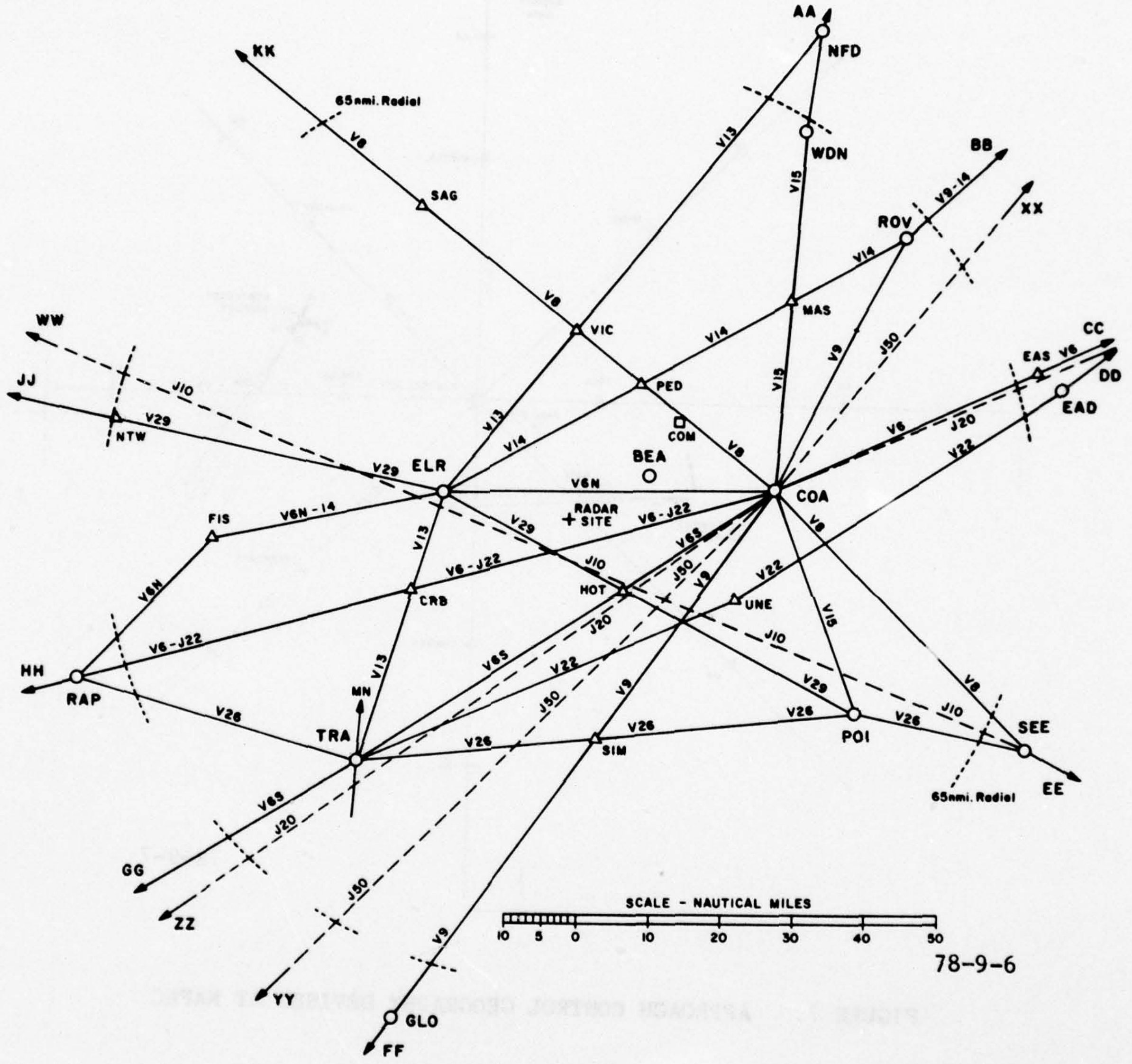


FIGURE 6. EN ROUTE SECTOR GEOGRAPHY DEVISED AT NAFEC

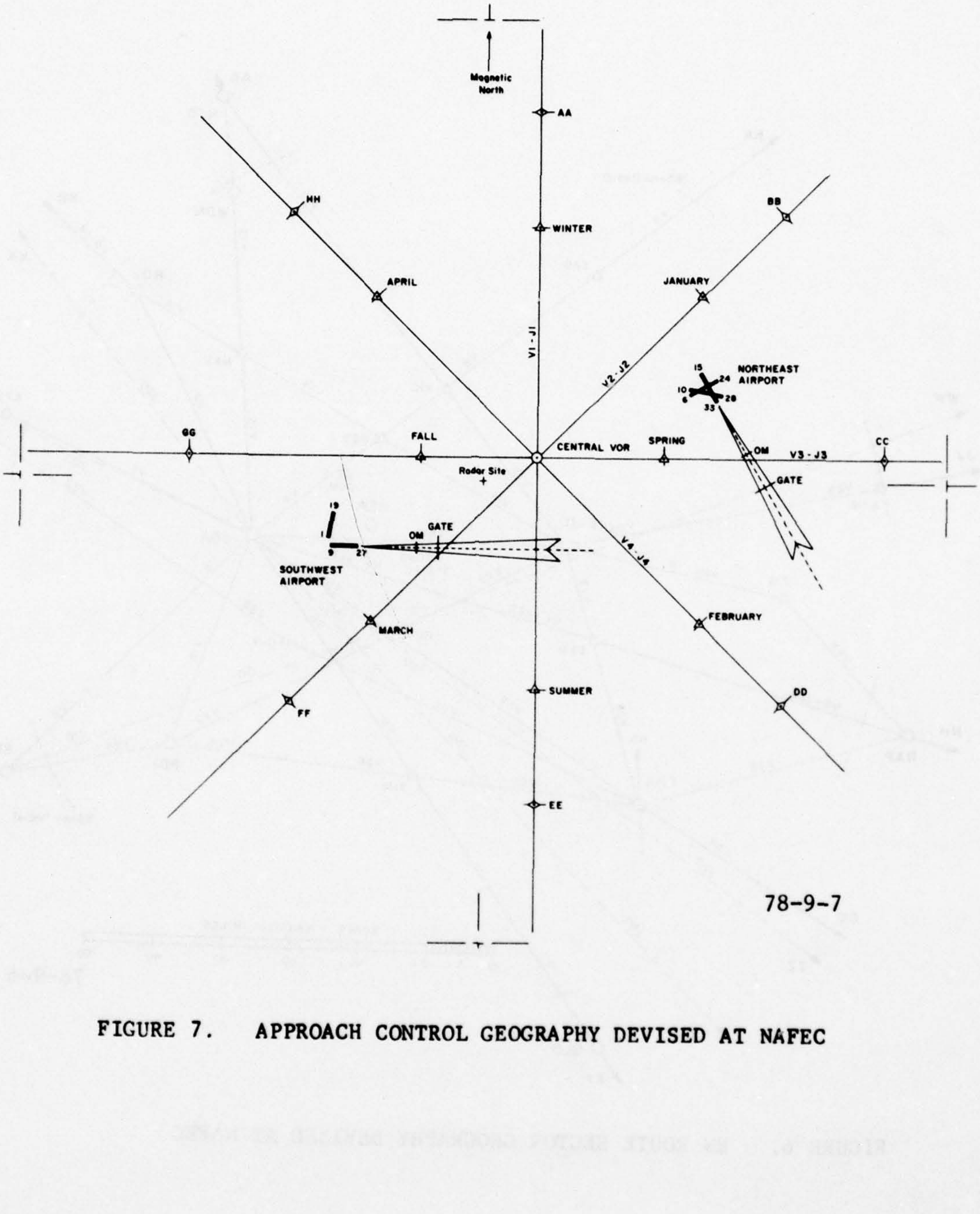


FIGURE 7. APPROACH CONTROL GEOGRAPHY DEVISED AT NAFEC

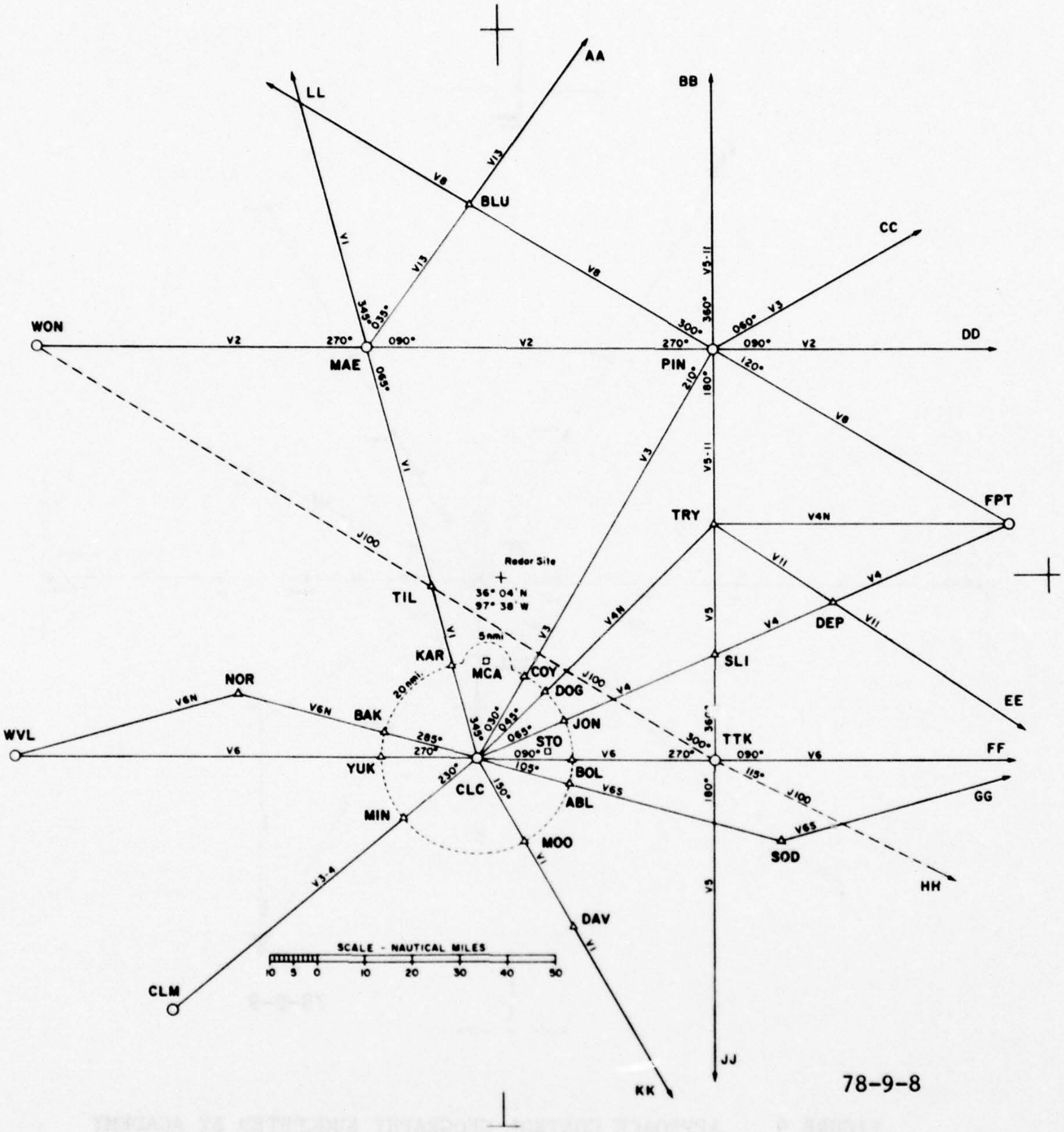
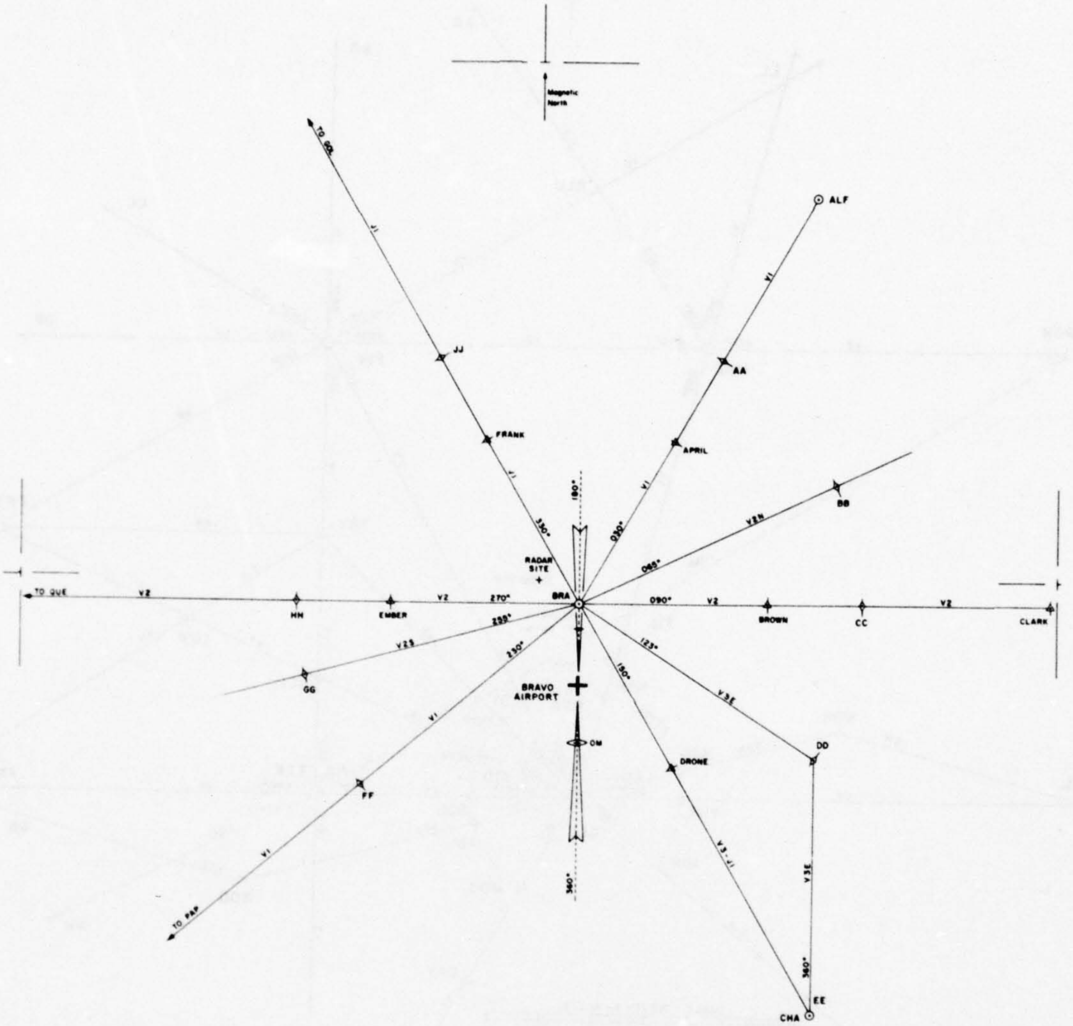


FIGURE 8. EN ROUTE SECTOR GEOGRAPHY SUGGESTED BY ACADEMY



78-9-9

FIGURE 9. APPROACH CONTROL GEOGRAPHY SUGGESTED BY ACADEMY

within the capability of a good, experienced controller. The actual level of difficulty to be optimally used for a given test group (journeymen or developmentals) would have to be the one determined on a tryout basis to be such as to yield a good score distribution.

PROBLEM LENGTH.

Problem length also requires careful consideration. Earlier limited tests (reference 8) confirmed that "two hours was a minimum run length for reliable data." If necessary, it is possible to split the tests into smaller segments and accumulate, or average, the results. Thirty-minute segmentation is built into the system. (See constant C-5 in appendix D.) Further experiments are necessary to establish the optimum run length.

PARALLEL FORMS.

Parallel forms of the tests are also required. The purpose of parallel forms is to make available different tests which are equal in difficulty. These would be used for retesting if required and also for preventing the population learning the substance of the tests. An early test (reference 7) examined one method of constructing parallel forms. This method was utilized in the development of parallel forms of the basic tests for initial experimentation.

TEST SUBJECTS

EARLY TESTS.

Some limited "PROBE" tests were previously accomplished at NAFEC to provide information and to guide the development of the controller performance measurement system (references 7 and 8). The sample of subjects available, however, was very small, and the data points were few. A much larger random sample of representative air traffic controllers is needed in order to obtain meaningful and reliable data to validate the concept and to understand the meaning of, or interpret, the scores.

SELECTING APPROPRIATE SUBJECTS.

Originally, it was assumed that normative data from two major groups would be required: Developmentals just graduated from radar training at the FAA Academy and journeymen controllers in the field. The first planned operational application of such tests was to be at the Academy; and therefore the group first requiring normative data was the Academy graduates. However, as the Academy does not presently provide radar training, the only source of a group of subjects with a relatively consistent (measurable) level of competence is the journeyman controller. It was, therefore, planned to start with this group.

The selection of the test subjects requires that a distinction be made between the basic experimental testing, for the development and verification of methodology, and testing for the purpose of accumulating a data base and normative data. Fundamental experimentation is necessary just to demonstrate basic methodological requirements such as measurement reliability, the required length of the test and development, and verification of comparable forms, as discussed later. It is doubtful that resistance to such experimentation, to gain knowledge for the good of the system, will be received from the field controller or his employee organizations, given certain nominal conditions. For instance, procedures can be developed to insure that the results obtained on a specific controller are not available for other uses and that all controller identity is destroyed after the experiments are completed. In fact, the only reason for ever needing to identify the subject would be to provide correlation between two separate test administrations, and this could be handled by codes rather than by names.

One question that needs to be answered is whether this type of information and verification could be obtained using journeymen controllers as subjects and be of any utility for the problem of testing developmentals at the FAA Academy. The assertion that there was any such utility was greeted by some as an intent to simply use the journeyman test for the developmentals without any change. On the face of it, of course, this would be unfair, since a developmental just out of the Academy could not possibly do as well as a journeyman with all that extra training and experience. Actually, it was the intent of the experimentation to provide assurance that ATC proficiency among journeymen could be measured meaningfully, reliably, and objectively. The verified methods and measures would then be used (with easier traffic samples) with Academy developmentals who have reached the point in their training where they began to function in the simulator in a true multiple aircraft traffic control manner. In addition, the test for journeyman level, having already been developed and its normative data accumulated, would then be ready for later use. The primary later use envisioned was that the journeyman level test could be used as the criterion against which to validate the predictions of the Academy level test, thus giving a valid way of determining the pass/fail level for the Academy simulation testing.

When it was determined that sufficient journeymen controllers could not be provided from the field, our plans reverted to the original desire to use developmentals. Without the Academy RTF in operation, however, it was necessary to consider use of those developmentals from the field at a level of training commensurate with that expected upon completion of the planned Academy curriculum. Obtaining such subjects from different facilities with various training approaches posed many problems. This approach, however, also has many advantages, as discussed in appendix E. The experimental design described there planned to use such developmentals, but it would apply equally to journeymen subjects.

NUMBER REQUIRED.

Initial plans called for a basic experimentation phase during which accumulation of a normative data base for journeymen controllers would be started. As such a base would be required eventually and could have many possible uses, a buildup of the journeyman normative data base was planned to continue following the basic experimentation. A normative data base buildup for Academy developmentals was also planned after the Academy RTF was in operation, providing basic radar training to developmentals.

To meet the customary professional criteria for establishing normative data, the number of controller subjects must be related to the total population (over 20,000 journeymen controllers) and to the number of measurements taken. About 600 journeymen controllers were estimated to be required to establish a normative data base. When funds were lost for the continuation of the MSL method of data collection, it was soon affirmed that the cost of the only reasonable alternative (bringing to NAFEC the large number of controllers needed to establish a normative data base) was indeed prohibitive. However, the appropriate strategy at this point would be to go to the basic experimentation described in appendix E, which has the limited objective of establishing the methodology and its verification with no aim of accumulating normative data. The minimum number of tests and subjects for this purpose is best determined after some data samples have been collected. Thus, that experiment is designed in modules of 30 subjects. Two or three such modules should be sufficient. Journeymen would be preferred as the subjects, because of the problem of obtaining developmentals at a consistent training level from various facilities, but developmentals would be acceptable.

EXPERIMENTATION REQUIREMENTS

GENERAL.

All of the laboratory and problem development efforts described in the preceding sections have been completed, except the MSL. In addition, system measurements which appear appropriately sensitive to system variations have been incorporated into the data collection capability of the ATCSF. (Appendix D contains a detailed list and explanation of these measures.) However, in order to be used effectively, certain characteristics and options must be demonstrably present in the finally developed system. Among these are distribution information, test-retest reliability data, and the availability of parallel forms (reference 12). The controller performance measurement concept is presently ready for the experimentation necessary to insure that such characteristics and options are present. Experimentation is also necessary to establish (on the basis of data on subjects), proper traffic density and minimum run length. Appendix E is a detailed description of keystone experiments prepared for this project by Computer Sciences Corporation to assure that the tests meet the necessary criteria, are properly sized, and that an appropriate grading system is available.

CONTENT VALIDITY.

Content validity is inherently present, in that the laboratory and problem structure embody the essence of the real-life radar control functions. Exposure of the test to a sample of current air traffic controllers would provide verification of this.

RELIABILITY EXPERIMENTATION.

Determining the dependability or repeatability of a test is customarily accomplished by repeating the test on the same group of persons and computing the correlation coefficient (product moment) between the two administrations. This is an absolutely essential process, and if a test does not meet this requirement (i.e., an adequately high coefficient of correlation), it would have to be abandoned. Previous work by Buckley et al. (reference 4) indicates that this requirement can be met.

PARALLEL FORMS EXPERIMENTATION.

Although initial limited tests have shown that adequate parallel forms can be prepared, assurance that a test is, in fact, a proper alternative for another test requires administration of the two tests to the same group of subjects and computation of the correlation coefficient between them. Various adjustments to the tests may be required to obtain an adequately high coefficient of correlation for use of the tests as parallel forms.

DATA COLLECTION AND INTERPRETATION.

The individual measures accumulated during a run of the tests (appendix D) are only meaningful in a comparative sense, not in an absolute sense. For example, for a given traffic sample, there is no answer available from anywhere as to what is an excessive number of delays, especially if the traffic sample is a rather heavy one. The number of delays is only meaningful for that traffic sample and in comparison to what delays other controller subjects incurred. For this reason, in order to be able to understand the meaning of, or interpret, the scores, it is necessary to accumulate the scores on a given traffic sample of a reasonably large and representative group of comparable traffic controllers.

There is no particular mystique to the process; it involves the collection of data on a large group and storage in such a way as to be capable of retrieval for comparative purposes.

In this connection it should also be mentioned that the interpretation process is not as simple as it might be, even given the normative data collection having been completed. The tests proposed here will yield multiple measures, for example, delays, delay time, number of aircraft handled, conflictions, etc. (appendix D). These measures are not equally important and are all interrelated, sometimes inversely (i.e., if one goes up, the other must go down). The point is that they should be interpreted in terms of the whole set of measures, as a profile of measures. The planned experimentation was designed to determine the best form of a grading system which would provide a single number index predictive of the subjects performance on a range of other ATC situations. The advice and consultation of Dr. Albert E. Beaton of the Department of Statistics, Princeton University, was utilized in the development of this experimental design. This experimentation, however, is only a first step designed to provide the minimum necessary assurance of the validity of the concept and to refine the methodology and grading system.

TEST APPLICATIONS

GENERAL.

The IDA report (reference 2) described a number of important applications of objective controller performance measurement tests of the type described herein. (Also found in appendix A.) Among them were validation and improvement of selection tests and procedures, evaluating the effectiveness of alternate methods of training, evaluating developmental and journeymen proficiency, determining the impact of new or proposed ATC equipment, and improving the technique for assigning controllers to facilities. Appendix C examines some of these uses in more detail and also describes some additional applications; the examination of models of controller workload and validation of mathematical models of controller functions. It is not our intention to repeat those discussions here, but the following recaps the applications discussed and adds some required further explanation regarding three applications: Academy pass/fail testing, utilization of the tests by management, and system test and evaluation. An adequate appreciation of the possibilities discussed, however, requires an understanding of the flexibility and variations of the basic test and analysis capabilities which are possible.

BASIC TESTING FLEXIBILITY.

The unique ATCSF capability to interface with the en route Simulation Support Facility (SSF) and the *Terminal Automation Test Facility* (TATF) *at* NAFEC opens the door to a wide variety of testing possibilities using variations of the basic test and analysis capability described here. For instance, control positions in the SSF or TATF could be used (instead of the controller positions in the ATCSF), thus providing complete operational realism and allowing introduction of additional measures gathered in the SSF or TATF. In fact, if desired, it is possible to interface the ATCSF with other external test facilities (as with the planned MSL) or with operational facilities. All of the explorations conducted on the MSL during this project indicated that it was technically feasible and simple to transmit simulated targets from the ATCSF to any location. In fact, use of central simulators, interfaced in this way, may some day supplant dynamic simulation training systems now used in the field facilities. In addition, the individual controller subjects could be replaced by full teams of controllers in either the ATCSF or other facility, if desired. Once the basic experimental design has been executed, the test technique, measures, and analysis can be applied in a number of possible variations for different purposes. It is important to note that the measures described in this report would remain essentially the same and would always be collected and analyzed in the ATCSF without impact on the other facility. Of course, additional subjective or objective measures of other controller activities, such as communications, keyboard actions, coordination activities, etc., or controller psychophysiological reactions, could be collected manually or automatically at the other facility as needed for the precise purposes involved. These could then be combined with the ATCSF measurements, if desired.

ACADEMY TESTING.

The technically best way (and all are agreed on this) to establish the validity of a pass/fail simulation testing program at the FAA Academy is to correlate performance scores on ATC simulation exercises at the Academy with performance scores on ATC simulation exercises at a later stage in the controller's career (longitudinal testing, also discussed in appendix E). Only then can a standard pass/fail criterion (grade) be established nonarbitrarily. The accumulation of a journeyman controller data base, therefore, is essential to this, as well as to other applications.

There is an important distinction to make here; one of the confusing factors in the situation is the two views it is possible to take of testing at the Academy. One view is that all testing at the Academy has as its almost exclusive main purpose, achievement testing, i.e., the determination of whether what the instructors have taught has been learned. The other view of pass/fail testing is that the most important purpose is to make a prediction at the Academy (and indeed, if possible, before that) as to who will eventually make it through the very expensive training and become a journeyman-level controller--presumably an adequate one. One's first thought will be that these objectives are complementary and not contradictory. However, confusion between them can cause considerable chaos. For, if the test is to be predictive, then an objective standard of success as a journeyman is required against which to measure and refine the pass/fail (predictive) test. Otherwise, only an understanding of the curriculum and the instructor's assessment of the students assimilation of the information taught is required.

An important point to be noted here also is the fact that the few weeks of Academy radar training is not intended to prepare a developmental for work as a radar controller when he arrives at his assigned facility. Several years of other activities and specialized facility-specific radar training are necessary before a developmental starts radar control of traffic. Therefore, the Academy's role would seem actually to be primarily an introduction to radar for the purpose of predicting the possibility of success. Actual achievement of radar control capability obviously is of lesser importance, as the information learned will be repeated, considerably expanded, and combined with other control experience before the capability is actually put to use several years later.

UTILIZATION BY MANAGEMENT.

In fact, both kinds of tests (achievement and predictive) are needed at the FAA Academy and in the field. The instructors or facilities need to have the freedom to develop the curriculum as they feel it is best to present the required information, and to establish tests they deem necessary to assure assimilation and application of that information. An important item frequently overlooked, however, is the need of training management for a means to assess the overall training effectiveness of the Academy, the region, the facility, and indeed of the whole system. Sets of common general tests, developed, maintained, and administered independently of the Academy or field facility could be used as an external and independent test, analogous to high school regents or college board scholastic aptitude tests (SAT's). Thus, they would provide training management with an overall quality control tool for training. Variations in the score distributions, resulting from these test applications over a period of time, or between facilities, would indicate either a change in the quality of the subjects (changes in selection or in previous test predictions) or in the quality of the training rendered. For this purpose, it would neither be necessary nor desirable to identify by name the specific controllers tested. Only the class or facility would be required to be identified. Information would be available for making specific management decisions relative to alternate training methods; order, type, and magnitude of training; improvements to the curriculum; setting standards for advancement or second career; etc.

Additionally, or alternatively, an aid for annual proficiency evaluation on an objective basis would also be available if the FAA and the employee organizations were interested. Every controller could be evaluated relative to a standard national score applicable to his particular level. This would not necessarily, or desirably, supplant the supervisory evaluation, but it could certainly supplement it and could be of great use to the individual supervisor. In addition, it would add objectivity, as distinct from subjectivity, to the process.

Central control of such a management tool is essential for security of test materials. The test methodology and system capabilities discribed here show its feasibility. Further research would be needed, however, before this application could be instituted.

SYSTEM TEST AND EVALUATION.

A recognized problem in the development of changes to the ATC system is the inability to obtain an accurate measure of the amount of improvement (or other impact) actually obtained (or expected) by the introduction of new software, equipment, or a whole new system. An accurate objective comparison between the systems (with and without the change) is possible with this testing methodology using the unique ATCSF capability for interfacing with other facilities described. The proposed hardware or software change could be introduced as the variable in a pair (or series) of tests, while the problem and subjects (individuals or teams) remain constant. Incorporating such tests in the test and evaluation of a prototype system, performed in NAFEC's unique simulation laboratories, would allow analysis of the real impact before a commitment is made to purchase and install large quantities of the change.

A clarification is necessary on the relevance of the keystone experiment (discussed earlier under Experimental Design and in appendix E) to the future of system tests and evaluations. Most system tests in the past have used the objective measures, such as confliction counts, delay times, and so on, which have been adapted for these tests. But they have always used small numbers of subjects, whether the subjects were individuals or control teams. Generally, the subjects were employees at the system test facility, and, while they were qualified controllers, they had often not controlled live traffic for some time. They were also thoroughly familiar with the test environment. The use of statistical experimental design techniques wherein the same controller teams (made up of these controllers) worked in all system variations being compared, in a random order sequence, was a skilled application of the tools at hand. But nevertheless, this leaves a good deal to be desired, both in the representativeness and the numbers of the subjects. Moreover, there was never the opportunity to discover the true shape of the distribution and the range and nature of individual differences among controllers or teams of controllers, thus leading to uncertainty about the sensitivity of the measures. This point is discussed in more detail in appendix F.

The point being made here is not that it will always be necessary to use very large samples of controllers in system tests (although representativeness should be handled more carefully), but that at least one experiment with a large sample is necessary in order to learn how large a sample is the necessary size. In addition, there is another benefit to be gained. It may well be that the major benefit of automation is not that it changes the level of performance of the system, in the sense that it changes the mean level, but rather that it affects the standard deviation or spread of individual differences, tending to reduce the variation among controllers. But the spread of individual differences is not known. It is at this point that personnel factors and system factors come together and become the same interest.

The performance measurement technique described herein also provides a capability not now available, for verifying and refining mathematical models designed to predict the effects of planned new systems before development. This is covered more completely in appendix C.

CONCLUSIONS

Due to early cancellation, the attempt to develop completely objective measures for controller performance in handling simulated ATC problems was not completed in this project. However, enough progress was made and certain insights were developed which make the recording of them for future use worthwhile. Among them are the following:

1. It is possible, and it has been accomplished, to program a computer to automatically collect all measures of ATC system performance customarily used in simulation exercises.

2. The application of such system measures is equally logical when testing various systems having the same or similar controller teams, or when testing various controller teams (or single individuals) utilizing the same hardware/ software systems.

3. Information is needed through experimentation in order to make fruitful use of the above. Experiments must be done to measure reliability, sensitivity, distributions, etc.

4. Such fundamental knowledge about system measurement is equally the keystone of further progress in the evaluation of systems, training, and personnel.

5. The major problem is the transport of controllers from the field sites to the simulator's location to take part in the exercises (since current field simulation devices do not incorporate sufficiently sophisticated software). Investigations made at NAFEC indicate that this can be overcome by transmission of simulator signals to the field.

6. The explicit experimental design suggested herein (a series of comparatively small experiments) is a keystone for all possible applications and a prerequisite for any additional experimentation required by specific users.

RECOMMENDATIONS

It is recommended that:

1. The small basic keystone experiments necessary to establish the reliability, sensitivity, score distribution, etc., of the performance measurement methodology described herein, be started as soon as possible. Additional experiments to meet the collective or precise needs of individual users should then follow as soon as subjects are made available.

2. Various possibilities be pursued for expansion of the special capability to transmit simulator signals to the field in a manner which allows field controllers to participate, from consoles at the field sites, in NAFEC's concept development, and test and evaluation experiments. With such a capability, a major problem (availability of current controllers for experiments in NAFEC's unique simulation laboratories) would be solved. Test subjects for data collection (i.e., obtaining adequate normative data) would also be available, in large quantities if desired. Operational impact of such a concept would be virtually nil, and costs would be reasonable (compared to transportation of subjects to NAFEC).

3. The possibility be investigated, utilizing the transmission capability described in 2 above, to obtain onsite data collection in support of various longitudinal studies relating to selection and training, and possibly for conducting proficiency development training.

REFERENCES

1. Anonymous, Engineering and Development Program Plan - Air Traffic Control Specialist Personnel Support, DOT, FAA, Office of Systems Engineering Management, National Technical Information Service, Springfield, Virginia 22151, Report No. FAA-ED-21-3, December 1975.

2. Henry, J. H., et al., Training of U.S. Air Traffic Controllers, Institute of Defense Analysis (IDA), Arlington, Virginia, Report R-206, January 1975.

3. Anonymous, Selection and Training of FAA Air Traffic Controllers, 12th Report, Committee on Government Operations, Union Calender No. 390, House Report No. 94-788, U.S. Government Printing Office, Washington, D.C., January 26, 1976.

4. Buckley, E. P., et al., A Comparative Analysis of Individual and System Performance Indices for the Air Traffic Control System, DOT, FAA, National Aviation Facilities Experimental Center (NAFEC), Atlantic City, New Jersey, 08405, Report No. NA-69-40, September 1969.

5. Corson, J. J., et al., The Career of the Air Traffic Controller - A Course of Action, Report of the Air Traffic Controller Career Committee to the Department of Transportation, National Technical Information Service, Springfield, Virginia 22151, NTIS No. AD-700 925, January 1970.

6. Woody, E. A., Director, Special Staff, AMN-20, NAFEC Research on Air Traffic System Performance Measurement, Letter to ARD-600, November 15, 1972.

7. Buckley, E. P., Development of a Performance Criterion for En Route Air Traffic Control Personnel Research Through Air Traffic Control Simulation: Experiment 1 - Parallel Form Development, DOT, FAA, National Aviation Facilities Experimental Center (NAFEC), Atlantic City, New Jersey 08405, Interim Report No. FAA-RD-75-186, February 1976.

8. Buckley, E. P. and Rood, R., CPM PROBE Experiment on Performance Information Feedback, DOT, FAA, National Aviation Facilities Experimental Center (NAFEC), Atlantic City, New Jersey 08405, NAFEC Letter Report No. NA-77-18-LR, April 5, 1977.

9. Bona, L.J., A Real-Time Simulation Facility for Evaluation Advance Concepts in Air Traffic Control, 1976 Summer Computer Simulation Conference Proceedings, Library of Congress Catalog Card No. 75-19746, page 989, July 1976.

10. Barnum, J. W., Deputy Secretary, Department of Transportation, ARTS-III Enhancements Program, Letter to the Federal Aviation Administrator, Washington, D.C. AOA No. 495, April 2, 1976.

11. Anonymous, Digital Simulation Facility Users Guide, Simulation and Analysis Division, Systems Development Branch, DOT, FAA, National Aviation Facilities Experimental Center (NAFEC), Atlantic City, New Jersey 08405, June 1975.

12. Davis, F. B., et al., Standards for Educational and Psychological Tests, American Psychological Association, Washington, D.C., 1974.

APPENDIX A

PERTINENT EXCERPTS RELATIVE TO CONTROLLER PERFORMANCE MEASUREMENTS

"TRAINING OF U.S. AIR TRAFFIC CONTROLLERS" (IDA REPORT R-206), JANUARY 1975

PAGE 27.

"Objective performance measures have important implications for the evaluation of alternative methods of training. They also have important applications, as mentioned above, in the selection of air traffic controllers, in the evaluation of controllers' performance on the job, in determining controllers' maximum useful workload for purposes of sector design, and for testing controllers' performance on proposed improvements to the air traffic system. Despite such value, objective performance measures are not in use at present, although their feasibility has been demonstrated in work at the National Aviation Facilities Experimental Center (NAFEC) (Buckley et al., 1969, 1972)."

PAGE 30.

"METHODS USED TO ESTABLISH COMPETENCE. Objective tests used for classroom or textbook subjects are thoroughly appropriate and relevant to the instructional material in the training program. However, determining the proficiency of performance at the ATC console (i.e., over-the-shoulder evaluations) is a subjective procedure which has been demonstrated experimentally to have limited reliability. The feasibility of objective performance measures has been demonstrated at NAFEC, but such measures are not in use at present.

"SELECTION. FAA studies show that selection procedures could be improved by incorporating performance tests that measure behaviors required of controllers at work. Further improvements are possible by assigning candidates to the en route, IFR, or VFR option on the basis of their test scores. Such improved selection procedures would be expected to reduce the number of candidate controllers who fail to complete their training."

PAGE 58.

"The FAA has supported and is supporting research that has direct application to training. However, the current level of support is so modest that it will take a long time before the needed results become available. Therefore, several research areas have been identified in this study as worthy of immediate attention:

"SELECTION. Little is known about how the criteria used during selection affect assignment to controller option or the quality of performance after completion of training. Improved selection procedures would be expected to reduce attrition during training and thereby save some of the expenses due to attrition.

"PERFORMANCE MEASURES. These are very important to the FAA, because together with other factors such as cost and flexibility, they provide the means needed to evaluate the ultimate effectiveness of alternative methods of training air traffic controllers. Performance measures are also needed for many other purposes of interest to the FAA, such as determining, for example, the controller's maximum useful workload, the distribution of traffic loads among sectors, the impact of new or proposed types of ATC equipment, and the significance of various tests and criteria for the selection of controllers. Such research should be focused on objective performance measures that could be employed in conjunction with the automated equipment at centers and terminals."

PAGE 153.

"The pioneering work of Buckley, O'Connor, and Beebe clearly demonstrates that it is feasible to measure the performance of an air traffic controller in an objective and reliable way. Here, some of the major results of their unusually comprehensive 'preliminary' investigation are noted. This work is still going on, and the current NAFEC simulator is improved over that used in the initial study."

"Buckley's current work at NAFEC is directed toward the development of a Controller Performance Measurement (CPM) test package which would include procedural instructions, scoring methods, and normative data (NAFEC Agreement No. 21-254, dated 22 June 1973). This is a research and development effort planned to continue over several years."

PAGE 158.

"Present procedures for evaluating the performance of controllers center around the over-the-shoulder evaluation, are subjective, and show low consistency on repetition, i.e., low reliability. FAA studies, which are still continuing at NAFEC, show that the performance of air traffic controllers can be measured in an objective way. The objective measures show a higher reliability than the subjective ones. For performance measures to be reliable, it is necessary to observe controllers performing on samples of air traffic carefully standardized for level of difficulty, i.e., density, complexity, and potential conflictions. For reliability, it also appears necessary to observe a controller's performance over reasonably long periods of time, so as to provide a measure which is the average of at least two 1-hour periods of observation. Objective performance data are needed to evaluate the effectiveness of various methods of training. They would also have great value in selecting controllers, establishing qualification standards, evaluating the proficiency of developmental and journeyman controllers, determining controller workloads at various levels of traffic, and thereby contributing to the design of en route and terminal sectors."

PAGE 159.

"Studies at NAFEC have shown that the performance of controllers can be measured in an objective manner. These measures show a higher reliability than the over-the-shoulder rating methods currently in use. Objective performance measures are needed to assess various methods of training controllers, the proficiency of controllers, and optimum workloads, and for similar applications concerned with the overall efficiency of the air traffic control system."

"TRAINING OF U.S. AIR TRAFFIC CONTROLLERS"
FOLLOW-UP ARTICLE IN INTERFACES MAGAZINE
VOLUME 7, NO.1, PT 2, NOVEMBER 1976

PAGE 51.

"Recognizing the importance of objective performance measures, the FAA initiated steps to have them applied nationally for improving and standardizing training, and for qualification and proficiency testing. To quote from the new FAA program, the result will be 'a nationally standardized facility training program for each option' (i.e., each type of job in air traffic control)."

"Improved appreciation of the value of measuring controller performance has resulted in the authorization of further research to improve methods of measurement. The performance of experienced controllers will be measured at operating sites in a mobile control center connected to the digital simulation facility at NAFEC which will provide standard air traffic samples and produce objective performance measures in near real time. The results will be used to establish objective performance norms for selecting and training controllers. It is too early to determine whether the FAA will fully implement the use of objective measures. But the fact of improved sensitivity to this issue coupled with the potential value of simulators leads to a belief that better evaluation of controller performance will eventually result."

"SELECTION AND TRAINING OF FAA AIR TRAFFIC CONTROLLERS,"
TWELFTH REPORT BY THE COMMITTEE ON GOVERNMENT OPERATIONS
JANUARY 1976

PAGE 7.

"Mr. Whitfield, in his prepared comments, defined the impact of the current attrition rate on the training budget of the FAA:

'Attrition costs are essentially in direct proportion to the number of controller hires. Based on current attrition experience (fiscal year 1973-74), about 23 percent of the developmental controllers hired in a fiscal year would resign or be separated before reaching journeyman level. An annual hiring intake of 1,800 controllers would, during the developmental period, result in : (a) the loss of about 400 developmental controllers equating to about 800 man-years; and (b) the loss of about $13.8 million invested in salaries and training.'"

"The IDA report stated that if trainee losses could be reduced by 50 percent, training cost reductions of approximately 16 percent for en route training and 11 percent for terminal training would result. While precision in estimating potential savings which would be realized from reduced attrition is not possible, it has been suggested that a reduction of 24 to 40 percent can be achieved once improved screening techniques have been implemented.

"By way of illustration, were the present attrition rate reduced by 25 percent, there would be almost $5 million in cost savings for the 1,550 individuals in the class that began their training in the Academy in 1975."

"A further technique to reduce attrition and insure, as much as is possible, optimal use of employee resources, would be to use previously discussed screening methods to help determine job assignments for each trainee. A discussion of the current assignment procedure appeared in the IDA report:

'Another possible route for improving selection, with its corollary effect on training, is to act on the fact that terminal and en route controllers perform somewhat different jobs. This is demonstrated in the task analyses performed for FAA by System Development Corp. (1971, 1972a-d, 1974a-f). On the basis of aptitude tests, training performance, and experimental performance ratings, Trites, Miller, and Cobb (1965) concluded that en route, terminal, and flight service station (FSS) personnel differ in the characteristics required for job performance. This is strongly confirmed with a much larger battery of experimental tests in the EPA (1972) study. This shows that the inclusion of new tests, including psychomoter ones, could increase accuracy of assignment between IFR, VFR, Center, and FSS from 25 percent (a random possibility of success when no special criteria are used in selecting personnel for these four options) to 58 percent. Within the IFR and Center options, accuracy levels of 75 to 80 percent could be achieved. Even better selection should become possible when Buckley's work at the National Aviation Facilities Experimental Center (NAFEC) produces useful objective performance measures for controllers on appropriately designed traffic samples.

'Improved selection procedures should reduce attrition in training and attrition due to inappropriate job assignments. Reduced attrition can, of course, produce significant reductions in the costs of training.'"

APPENDIX B

AAD-1 LETTER TO AED-1, ASSISTANCE REQUIRED IN THE DEVELOPMENT OF ATCS TRAINING, DATED JULY 14, 1975

DEPARTMENT OF TRANSPORTATION
FEDERAL AVIATION ADMINISTRATION

WASHINGTON, D.C. 20591

DATE: JUL 14 1975

IN REPLY REFER TO:

SUBJECT: Assistance Required in the Development of ATCS Training

FROM: Associate Administrator for Administration, AAD-1

TO: Associate Administrator for Engineering and Development, AED-1

Recent investigations and evaluations have revealed a substantial inadequacy in the Air Traffic Controller Training Program. To correct the identified deficiencies, the Office of Personnel and Training, in concert with the Air Traffic Service, is presently engaged in an effort to develop a revised terminal/enroute training program.

During the development effort, it became apparent that in order to assure successful implementation of the program and meet the stated objectives, additional resources and development actions would be required.

In response to the above and to a need suggested in the IDA study, you established an AED working group on ATCS training which has been exploring areas in which your office could most actively support Personnel and Training in their effort. As a result of their meetings, the group has determined that the areas in which your assistance would provide the most immediate benefit are:

1. Academy Simulation

    a. Development of engineering specifications for Academy radar simulation according to the attached training requirements and schedule.

    b. Providing technical assistance to the Aeronautical Center, which I believe should be the Requiring Activity, in their procurement of the simulation system for the Academy.

2. Identification and development of improvements to field radar simulation which should be completed by January 1, 1976, to provide lead time for FY-78 budget programming. This should include as a minimum.

    a. functional enhancements;
    b. scenario development, tape input and adaption;
    c. engineering specifications for pilot consoles;
    d. future enhancements necessary to provide complete and total simulation of the operational environment.

3. Immediate assistance in the completion of CODE development.

4. Coordination with AAM, CAMI, TSC and NAFEC in the continued development of Controller Performance Measurements (CPM) which would accurately identify controller progress during the various phases of his/her training. The action should be completed by January 1, 1976, as stated in the proposed Air Traffic Controller Training Program completion schedule. (Copy attached).

5. Development of a capability for simulation training at the ARTS II facilities, VFR Towers and Non-NAS centers. In order to prepare this for submission in the FY-78 Budget, I suggest a completion date of January 1, 1976.

While seeking your immediate and active support in the areas enumerated above, I in no way mean to infer that this is the total support you could render. For instance, immediate and on-going expansion of coordination and communication with the Office of Personnel and Training regarding those E&D programs which have a direct impact on operational or maintenance training would be most beneficial. Some areas which should be considered are your developmental and research efforts concerning FSS, DABS, MLS/RNAV/M&S, ICS and WVAS. To accomplish this, I suggest retention of the E&D Working Group as a permanent interoffice activity. Should you, or the working group, determine other methods of assistance, the arrangement for action upon them would be greatly appreciated.

Charles E Weithoner
CHARLES E. WEITHONER

Enclosure

APPENDIX C

SYSTEM PERFORMANCE MEASUREMENTS AND INDIVIDUAL PERFORMANCE MEASUREMENTS, A PAPER PRESENTED AT THE SECOND INTERNATIONAL LEARNING TECHNOLOGY CONGRESS AND EXPOSITION ON APPLIED LEARNING TECHNOLOGY, FEBRUARY 1978

SYSTEM PERFORMANCE MEASUREMENTS AND INDIVIDUAL PERFORMANCE MEASUREMENTS

E. P. Buckley and K. W. House
Department of Transportation
Federal Aviation Administration
National Aviation Facilities Experimental Center

ABSTRACT

The mission of the air traffic control system has been stated to be the safe and expeditious control of aircraft traffic. This system mission statement can be analyzed to produce specific system performance measures. This analysis has been done and specific measures of aircraft safety and expeditious movement have been programmed into the digital simulator presently in operation at the FAA's experimental center in Atlantic City, New Jersey. The simulator presents real-time dynamic radar displays to teams of air traffic controllers working with proposed newly-designed air traffic control systems. The systems are comparatively evaluated in terms of the measures encapsulating the system's mission. This paper explains how this set of measures is being adapted to the measurement of the proficiency of the individual air traffic controller. This process will enable acquisition of information of value for the improvement of the technology of air traffic control system design evaluation.

INTRODUCTION

Clearly fundamental to all effective training evaluations, the task of describing the job and then evaluating the performance of the individual is quite difficult. It is particularly difficult in air traffic control and this paper will describe an approach to it which is believed to have some generality to other fields.

The job of the air traffic controller is particularly difficult to work with because it simply will not yield to the book solution of analyzing the tasks the control specialist performs. It is just about impossible to specify the tasks of the air traffic control specialist in such a way that the appropriate and meaningful measurements become apparent. Focusing on the tasks seems to lead to rather sterile lists of how long it takes to perform, in seconds, various observable movements. The reason why this is sterile is that the critical tasks are intellectual and so not observable as they happen. The observable tasks remain a minor aspect of the job and are of dubious relevance to the core of the work.

Interest in describing and objectively measuring and evaluating the task of the air traffic controller arises from many needs. In the first place, it is important to know the level of proficiency which has been reached at the end of each stage of training. Secondly, the average of such results can be used to evaluate alternative training programs or individual instructors. Criterion measures of performance are also needed in order that selection tests and procedures can be validated. There is also a need to periodically evaluate the skill levels of persons who have long since completed training in order to detect signs of needed refresher training.

Other methodological quests exist which do not particularly concern the skill of the individual air traffic controller. These are concerned with the evaluation of system performance for the sake of knowing whether changes in the hardware or the software of the system are needed to make its functioning more effective. In order to make any progress in system design, it is necessary to be able to evaluate the functioning of the system in various conceivable configurations and the terms of the evaluation must be independent of the particular design involved.

TWO COMMUNITIES

We have thus far delineated some groups, or communities, interested in performance measurements for the air traffic control system and its human operators. The question we mean to examine here is one concerning whether or not there might be some symbiotic relationships possible, and desirable, between these communities.

The two major communities involved may be described as the personnel and training community and the system design and engineering community. It is possible to very briefly review the approaches these two communities have taken to their evaluation problems. The personnel and training community has expended a lot of energy on individual controller evaluation. The history is long and intricate. The history is reviewed in great completeness in a very scholarly appendix to a recent report by the Institute for Defense Analyses (1).

The Personnel Community

The main tendency of the personnel community has been to use subjective ratings and to concentrate on individual techniques within the job such as the manner of communication, phraseology, and so on. Overall supervisor and instructor ratings have been used a great deal. The rating is usually done while observing the controller at work at his usual position. There has been a great step forward recently which was contributed by the Systems Development Corporation (SDC) in the shape of a rating which is carefully constructed to focus on objective kinds of things during the "over the shoulder" rating. Prior to the SDC form, the relibility estimates for ratings were around .40; SDC's seems to be, according to one report (1), about .65.

The System Evaluation Community

The approach taken by the system design and analysis community has, on the other hand, been objective but a little skimpy on concentrated methodological studies. It has concentrated on the technology of simulation and simulators and has improved these greatly. The simulators are now completely digital and wonderfully capable of flexible adaptation. By simulation here is meant, specifically, a real-time (as opposed to fast-time) representation of the task under study. Such a technique is more familiar in the case of aircraft cockpit simulators representing the cockpit and the dimensions of the pilot's task. Real-time man-machine system simulation has been used quite widely for system studies in many systems including aircraft and submarines. Perhaps the first major project using real-time simulation in system study was Project Cadillac (2) just after WWII which involved the simulation of an airborne air defense radar information processing center. A general survey of system studies using simulation can be found in the excellent book "Man-Machine System Experiments" by Parsons (3).

Since the primary information of use to the air traffic controller comes to view in the form of radar signals, the use of simulated radar signals provides a rather realistic simulation of the working environment of the air traffic controller. The Civil Aeronautics Administration (CAA), the forerunner of the FAA, had an impressive radar simulator operating at its Technical Development Center in Indianapolis in about 1950. From 1950 to 1958, approximately 50 technical reports were prepared from the CAA simulation work. None of the experiments had to do with the use of simulation in training, but rather constituted system design and procedure studies. When the CAA became the FAA, NAFEC took up this work of examing air traffic control systems and procedures by means of real-time simulation. Between its founding in 1958 and today, NAFEC has worked on many different types of problems using simulation.

Two topics, the exploration of optimal locations for new airports all over the world and the trying out and evaluation of new automation systems have been the main work of the NAFEC simulation laboratories. Most of the simulations collected objective measures of system performance. These objective measures focused on the comparative ability of the system to perform its mission under the various alternative configurations of hardware, software and geography that were being considered.

Community Characteristics

The approaches taken by the two communities, then, have manifested different characteristics and have developed different skills and acquired different experience backgrounds. FAA personnel research studies, conducted usually by the Civil Aeromedical Institute of the FAA, have used subjective performance estimates, usually done by supervisors and sometimes by instructors and training officers. The major purpose of estimates of proficiency in the research area has been for the validation of selection tests. These studies have used admirably large samples of controllers in recognition of the necessary precautions for sampling and statistical data interpretations. The measures have, of course, been readily interpretable in terms of goodness and poorness on the rating scale.

The engineering studies, on the other hand, while marked by the possession of objective measurement techniques, quite often gave insufficient consideration of the probable individual differences among representatively sampled controllers. In many of the experiments, statistical analyses were not done and subjective interpretation of the objective data was made. In others, statistical techniques were utilized but the samples of subjects were of extremely small sizes. In an engineering way, the controllers as such were regarded as a non-variable in the experiment; and it was frequently tacitly assumed that, as long as qualified controllers were used, they were sufficiently representative.

This paper is intended to be a discussion of how these two traditions can be blended. The discussion is not only relevant, it is hoped, for those who are interested in the air traffic control system. It is also a general paradigm discussion of the process of developing criteria in a manner as to overcome the deficiencies in both traditions and provide basic information simultaneously useful to system manning and system design.

THE SYSTEM ENVIRONMENT AND MISSION

Environment Analysis

The job environment in which the air traffic controller works is dominated by the radar scope. It is through this that he receives the primary information with which he operates. Other information is received through voice communication with the pilots of the objects (aircraft) he is controlling. Some information also comes to him through the computer and he also has to enter new information into the computer. The controller almost always works as a member of a 2 to 4 person team. The roles of these people are defined; but there is an immense amount of overlap and exchange of tasks depending on the situation. The central person, though, is always the radar controller (the R controller), who makes the basic control decisions. It is important to note that the other members of the team are either already qualified R controllers, but not working as such on that day, or are assistant controllers who hope and intend to be fully qualified R controllers some day. In addition to the centrally important decision making involved, there are, of course, other tasks. The pilots must be informed and consulted about the control decisions. Their requests must be received and responded to. The response acts must be done; but it is vital to note, that while sometimes the response is a simple giving of information, it is more frequently and more importantly the communication of a decision. But there are customs and conventions regarding the phraseology of communication which must be attended to. Similarly, there are techniques involved in entering the decisions which have been made, and the results thereof, into the computer; and there is a certain amount of skill in knowing how to make the entries in the form acceptable to the computer. There is also a task which is a form of bookkeeping which involves recording data on paper slips called flight strips.

The members of the control team other than the radar controller are largely involved with these auxiliary duties. The radar controller is the decision maker. If one does not give the team different equipment, then their various abilities come into play in determining the success of the air traffic control system in accomplishing its mission.

Sectors

The U. S. airspace is divided into what we can call sectors, manned by sector teams. There are in the continental U. S. a large number radar control sectors, approximately 500 of which are terminal sectors and approximately 700 of which are enroute sectors.

System Mission

The mission of the air traffic control system is clearly defined. It is the safe and expeditious movement of air traffic. In this context, the safe movement of air traffic involves aircraft getting from place to place on the ground (runways) and in the air without any collisions. The expeditiousness aspect of the mission involves the minimizing of what the airlines call "air traffic delays," i. e., delays to individual aircraft caused by the cumbersomeness of the traffic control procedure or the overall volume of air traffic in relation of the control system's ability to smoothly handle it.

That, then, is the mission of the system. It is possible to improve the ability of the system to fulfill its mission by improving the skill level of the people in the system and/or by giving them new, better tools to work with in doing the job, i. e., performing the mission. Radar itself was introduced as such a tool for helping perform the mission.

The "customers" of the system, the commercial airlines, the military, and the general-aviation pilots, would like the system's performance of its mission improved. But they are not terribly concerned with whether that improvement is brought about by improving the system's equipment, by improving the teamwork between man and machine, by improving the teamwork among the people who are the team members in the system, or by the improvement of the skill levels of the individuals who make up the teams. No matter how it is done, they, the customers, will be able to tell if and when it has been done, if they suffer less air traffic delays, and if the number of "near misses" reported goes down (assuming that the amount of traffic handled remains the same). Of course, the amount of traffic handled is constantly going up as the years go by. In fact, it somewhat shifts from day to day, from hour to hour, and from minute to minute. It certainly shifts from sector to sector.

MISSION AND MEASUREMENT

Situational Elements

There are then many ATC teams and almost all are composed of two or more people. All of these people have been trained and are being trained. They were also once selected to do this job and indeed also themselves chose to do this job. They are evaluated periodically. The evaluations are subjective and given by the supervisors after "over the shoulder" observations. They are concerned with how the person does at the one or few sectors at which he or she usually works. The evaluations cannot, strictly speaking, be compared with each other. Not only are the evaluators quite different, but the sectors are, to some unquantified extent, different. Some sectors are quite busy, others are not as busy, and some have more complicated route structures than others.

Such very place - specific evaluations might be all that we can do for personnel evaluation, but it would be desirable for the evaluation of training program effectiveness and for the validation of selection systems to have a measure of the effectiveness of teams and individuals which possessed comparability from sector to sector. It would also be best if such a measure was completely, or largely, objective and had little or no opinion in it.

But the performance of the air traffic control system, represented at the sector level, can be measured. Because most of the crucial tasks of the control team are truly describable as decision making, they are hard to observe and measure. However, the effects on the sector system of a set or sequence of such decisions can be measured in terms of the accepted criteria for the system, namely safety and expeditionsness. Skillful decisions are those which result in higher safety and expeditiousness. But higher safety and expeditiousness than what? Higher safety and expeditiousness than might result when the decisions are made by other people in the identical situation. But how can the identical situation be reproduced? By means of simulation. The same maps, the same aircraft the same schedules can be produced repeatedly and the same situation requiring sets of successive decisions can be presented to many teams. Then the results obtained by many teams in their handling of the identical situation can be compared. This comparison can be in terms of how well the several teams did in meeting the system mission when faced with the same traffic situation. This relative, not absolute, comparison is all that is possible, but it is meaningful in terms of system mission accomplishment in the identical situation.

If we regard the system as a unity made up of the team of people and the assemblage of equipment, and if the systems analysts have figured out quantitative statements (measures) of the system's effectiveness in attaining its mission of safety and speed, then these measures should be useful in another way. What the system designers and evaluators do is to keep the team the same and change the tools they are given (hardware, software) and then take their system mission performance measures. The same measures should be able to express the attainment of the system mission if the teams of people operating the system are changed and the hardware/software system is held the same.

There are, in other words, four variables in the situation: the system hardware/software configuration; the team of personnel who work with the hardware and software; the traffic situation; and the outcome variable, the measured system performance. If both the team and the system hardware/software configuration are kept constant and simulation runs are made, the runs can be repetitions or replications. That is, the traffic set, which is the stimulus the system (team and hardware) has to deal with, can be exactly duplicated in simulation on various occasions. (This is not the case in the real world system). So, any of the independent variables can be changed; the system, the team, or the traffic sector map complex. The ones which are changed and the ones which are kept the same will depend on the investigator's purpose. The advantage of the simulator is that one can choose which of the three independent variables it is appropriate to vary and observe the effect on the dependent variable, the index of system mission accomplishment. In short, if one wishes to compare teams, or hardware/software configurations, or traffic sets, in terms of the effect these changes have, on the system's mission accomplishment, this can be done.

The question recurs about the terms in which we will describe the accomplishment of the mission. It is not inconceivable that we could have skilled observers watch the session and have them report at the end of the session "GOOD, FAIR, or POOR" depending on how well or poorly the system-team combination appeared to accomplish the system's mission of safe and expeditious traffic movement. In making their judgement of "GOOD, FAIR, or POOR" they would, of course, have to consider the particular set or level of traffic the system-team combination was up against that time. They would also have to use, as their frame of reference, what they seemed to remember about other teams handling the same or similar traffic levels. Their judgements would be, consciously or unconsciously, in comparative terms, i.e., in relation to other teams and systems they had seen functioning in the same or similar traffic situation insofar as they could remember it.

Measurement Types

It is at this juncture that a lesson can be learned from the system engineers and analysts. They have a set of measures of system performance, which quantitatively express how well the system performs its mission. They use these to compare various systems configurations. Perhaps the same measures could be used to determine how well a particular team performs the mission.

We should now look at the measures that have customarily been used in system test and evaluation and see how it is that they represent quantitative statements of the air traffic control system's mission accomplishment.

Of course, the major purpose of the air traffic control system is to prevent collisions between aircraft and this is so important that the word collision isn't even used. The word used is confliction which refers to the event of two aircraft having violated the FAA's separation standard. The separation standard varies with different circumstances but a typical one is 3-miles horizontal and one thousand feet vertical. This means it is against regulations for the controller to allow two aircraft to come within those dimensions of each other at the same time. They can be within the 3-mile distance, however, if they are sufficiently separated in altitude, and so on. In the basic case, though, a confliction is the primary possible failure of the ATC system, and a controller should never let it happen.

The other major aspect of the air traffic control system's mission has to do with the expeditiousness with which the safely proceeding aircraft move. Aircraft may be delayed by the controller for reasons of safety, but his ideal is to keep the traffic moving as rapidly as it is capable of moving. There are three ways in which such delays occur. One is by simply not allowing the aircraft to take off at the scheduled time. Another is to put the aircraft in a hold, which is a circle or a racetrack pattern to be flown in the sky. Finally, the controller may indirectly delay the aircraft by refusing to accept an aircraft from the adjacent sector which wishes to hand the aircraft off to him. Closely related to the matter of expeditiousness is the matter of path changes. In guiding the aircraft, the controller attempts to keep to a minimum the path changes, such as altitude and heading changes, he requires of each aircraft. In sum, the controller's objective is to use tactical decision making in such a way as to guide the aircraft safely through the sky in the shortest possible time with a minimum of vacillating changes.

The NAFEC computer has been laboriously and successfully programmed to count and measure the concrete manifestations of such occurrences and states as listed above. For example, it records any moment at which any aircraft are within three (or x) miles and, simultaneously, a thousand (or x) feet of each other. It also counts instances in which aircraft are held in the air, held before takeoff or not accepted at hand off time. The basic measures (similarly derived from the mission of the air traffic control system) which the NAFEC computer automatically records, and their detailed definitions, are listed in a forthcoming NAFEC report (6). Additional measures are compounded of these simple measures, such as the ratio of the delayed aircraft to the number in the traffic sample, and these are also explained in detail in the same report (6). One large and two small experiments using these kinds of measures to look at individual controller performance in simulation have been completed (4) (5).

Multiple Measurements

The multivariate nature of the dependent variables is of great importance. Clearly the two most important aspects of the system performance criterion, safety and expeditiousness, are to some extent contradictory in the sense that a balance must be achieved between them. It is in this sense that the job of the traffic controller is an intellectual and executive function. Maintaining this balance at its optimum is his controlling role. The major measures are conflections and delays. Others, such as the number of aircraft handled and the number of altitude changes and vectors given to the aircraft to execute, also reflect the mission. The optimum balance ought to be seen as a pattern of these dependent variables.

One could look at the pattern, as a pattern, but this can just as well be accomplished through multivariate analysis. This is an improvement over the previously customary approach to system test and evaluation of analyzing one dependent variable at a time. But the decision to use multivariate analysis is not in itself an answer to the problem of considering the optimal balance. Instead, an experiment should be done which would determine a set of weights to be applied to the various aspects of the criterion represented by the multiple measures. One could "capture" the feeling of some expert traffic controllers about the optimum balance among the measures by having them observe a variety of experimental simulation runs which were also scored by the objective scoring system described above. From this data an analysis of the weights they customarily fundamentally utilize could be made by means of multivariate techniques. The proposed experimental design is discussed in detail in a NAFEC report (6).

IMPLICATIONS

A Hypothetical Experiment

Assume for the moment that the experiment has been completed and that it yielded a set of weights for the various dependent measures such as to yield a single index composite number indicating the relative goodness of system mission performance. Let us now visualize another but far more complex experiment. Admittedly, this experiment is impractical and indeed impossible, but its consideration will yield us a way of looking at the matter of system experimentation in a unified manner.

Let us design a very large air traffic control system simulation experiment. Let us suppose we could obtain the cooperation of a large group of controller teams, a large variety of sector maps and associated traffic samples of various densities of traffic and a sample of various conceivable hardware/software system configurations. Combinations of all of the independent variables could be exercised in simulation runs and their performance scored in terms of the system mission performance index. The results of the experiment could be expressed in terms of the variance in the system performance index attributable to the various sources.

Some of the resulting data would affect planning. For instance, suppose the proportion of variance in system effectiveness attributable to individual differences is markedly greater than that attributable to the various levels of automation? Taken at face value, at least, that might seem to indicate that all system engineering should be soft-pedaled in favor of concentrated efforts in training and selection of controllers. It might be of course, that the opposite was the outcome, but the most likely outcome is a complex interaction. But the course of such an interaction could guide plans and resource allocating. Another question would concern the important question of whether the variation among qualified controller teams was very large or very small in terms of air traffic control system mission achievement. The relative difficulty of various types of sectors and the form of the impact of traffic densities would be described.

By visualizing this huge and non-do-able experiment, we can more easily conceptualize a vital program of experimentation which has to do with many topics important to air traffic control as a whole, to both the system engineering people and the personnel and training people, whose interests are in fact inseparable, because they meet at the human factor. Obviously, a third group's interests come into play right here and that group is the technologists who need a great many of the basic methodological facts and distributions of data which would come from such a program before they can attack some of the specific problems they have been, and will be, asked to work on.

Applications

Following is a list of the areas of interest to air traffic control which would benefit, and a brief explanation of each. Not to be forgotten is the application of the generality of this approach to systems other than the air traffic control system.

1. Air traffic control personnel selection and training.

The same experimental methods and techniques which are useful for system test can be used to test individual controllers. In the experiment described above control sector teams were assumed to be the subject, or experimental unit, for analysis. If, however, the individual controller is substituted for the team, and the system remains the same, it is apparent that the system test comes to be a test of the individual controller's ability to accomplish the mission of the system by his decision-making, all other things being equal. Different traffic samples and geographies for various groups of controllers would be needed, but the fundamental methodology and measures would be generalizable and adaptable to various groups and uses. Data already exists showing that the system measurements are both reliable (repeatable) for, and relevant to, the examination of individual proficiency (Buckley O'Connor and Beebe, 1969) (4). A more recent report (Buckley, et al), also discusses a simulation using "one-controller, one-sector systems" (5). The central experiment necessary to develop the methodology basic to such uses is given as the experiment described in a forthcoming report (6).

2. The examination of models of controller workload.

There is a need for a mathematical expression of the workload or relative difficulty imposed by various geographic and traffic situations. Such mathematical expressions have been developed by Arad (7), and Ratner (8). They are generally based on map and traffic parameters such as the number of intersections in the map and the number of aircraft in the traffic sample and so on. This work is very important for the quantitative and a priori definition of sector difficulty. Such a definition is needed for sector redesign and also to enable the other areas to proceed more smoothly. For example, in the personnel area of experimentation, it will be necessary to express scores as norms for a given traffic sample, until a mathematical statement of sector geography and traffic sample comparability can be developed.

But as important as such mathematical expressions for situation difficulty level are, they need, once composed, to be validated and improved, if possible. The approach by which they can be validated, given the measurement system discussed herein, is rather simple. A series of maps for sectors and appropriate traffic samples can be composed and then characterised by the mathematical models in question. Then those problem situations could be administered to a resonable number of traffic control specialists. The level of difficulty gradient of the map-traffic sample combinations could be compared to the predicted order given by the mathematical models. This could be done until the mathematical models were sufficiently correct. This would enable indexing traffic sample and sector relative difficulty.

3. Validation of mathematical models of controller functioning.

The RECEP model by the Stanford Research Institute (8) is also a mathematical model of how the air traffic controller functions in his job. The model has been utilized to make estimates of the effects of future automation steps, such as the addition to the system of electronic flight strips, and to make statements of the effect of this on controller capacity.

The model is in need of empirical validation in order that such statements be dependable. The predictions of the model, as to controller capacity in given sector-traffic situations, could be verified by comparing predictions to results of simulation experimentation, just as in the case of the sector difficulty models.

4. Improvement of System Test and Evaluation Methodology.

As the air traffic control system is further developed it will be advisable and indeed necessary to determine which directions automation should take. Many variants of automation should be tried out and tested for their effect on the system's accomplishment of its mission. The measurement system described here is capable of doing this in an objective manner and also in a manner which will remain completely relevant despite the continuing changes in the system design. Since this measuring system is derived from the system mission and not the tasks within the system, it will remain timely as long as the purpose of the system remains the same. But a simulator and set of measures collected by the simulator are not enough to supply correct evaluations of levels of automation. It is necessary to know the shapes of the distributions of such measures in order to estimate the number of runs needed for system evaluation. It is also necessary to examine the sensitivity of the various measures and to see which measures are so highly correlated that they overlap. Some basic experiments to discover these facts will be necessary to prepare for future crucial system tests. The most efficient way to get this data is through experimentation runs with one-controller, one-sector systems.

SUMMARY

In summary, this paper has reviewed the mission of the air traffic control system and pointed out that certain objective and quantitative measures of the accomplishment of that mission have been developed for a long time for use in air traffic control system simulation. The paper points out that these measures can equally well be used to evaluate personnel teams or individuals, if the system is held constant, or to evaluate systems, if the personnel are held constant. To enable such uses in a skillful way, the same basic course of preparatory experiments is needed to yield information on measure sensitivity, reliability, and generalizability. The major point of the paper is to show how the development of performance criteria through basic experimentation can simultaneously improve methods for the evaluation of post-training performance and for system evaluation.

ABOUT THE AUTHORS

Dr. Edward P. Buckley is an Engineering Psychologist at the National Aviation Facilities Experimental Center (NAFEC) of the FAA. He received the Ph. D. in Experimental Psychology in 1954, with specialization in perception. He worked at the Naval Research Laboratories in air defense system studies and at the Engineering Psychology Laboratories of the Franklin Institute of Philadelphia on problems of decision making and air traffic control system design and evaluation. He came to the FAA in 1962, and has been concerned with air traffic control system evaluation methodology, air traffic control specialist aptitude testing, and all aspects of human factors in air traffic control. He is a member of the Human Factors Society, the Sigma Xi, and is a fellow of the Society of Engineering Psychologists of the American Psychological Association. He has made over 20 presentations and reports on human factors in air traffic control.

Mr. Kenneth W. House is Program Manager of the Air Traffic Control Specialist Personnel Support program at the FAA's National Aviation Facilities Experimental Center (NAFEC). The program was established to provide engineering and human factors research and development support, as requested by the FAA Office of Personnel and Training and the Systems Research and Development Service. He is himself an air traffic control specialist having acquired operational experience in all three traffic control fields; terminal control, enroute radar control, and flight service station operation. Prior to his present assignment, he has served in management at NAFEC as a member of the Engineering Management Staff and also as the lead FAA Systems Analyst coordinating enroute computer program development for the automated National Airspace System.

REFERENCES

(1) Henry, J. H., et al, Training of US Air Traffic Controllers, Institute for Defense Analysis, Arlington, Virginia, Report R-206, January 1975 (AD/S-006603).

(2) NYU CADILLAC Staff, Project CADILLAC Summary Report: Recommendations for Operating Procedures and Personnel Allocation in the CIC Compartment of the WV-2 Aircraft, Technical Report SPECDEVCEN 279-3-23, New York University, New York, 1956.

(3) Parsons, H. M., Man-Machine System Experiments, Johns Hopkins Press, Baltimore, MD, 1972.

(4) Buckley, E. P.; O'Connor, W. F.; and Beebe T., A Comparative Analysis of Individual and System Performance Indices for the Air Traffic Control System. DOT, FAA, National Aviation Facilities Experimental Center (NAFEC), Atlantic City, NJ 08405, Report NA-69-40, September, 1969.

(5) Buckley, E. P. et al, Development of a Performance Criterion for Enroute Air Traffic Control Personnel Research Through Air Traffic Control Simulation: Experiment I-Parallel Form Development, DOT, FAA, National Aviation Facilities Experimental Center (NAFEC), Atlantic City, NJ 08405, Interim Report FAA-RD 75-186, February 1976.

(6) Buckley, E. P.; House, K. W.; Rood, R., Development of a Performance Criterion for Air Traffic Control Personnel Research Through Air Traffic Control Simulation, DOT, FAA, National Aviation Facilities Experimental Center (NAFEC), Atlantic City, NJ 08405, Final Report, forthcoming.

(7) Arad, B. A., "The Control Load and Sector Design," The Controller, Vol. 3 pp. 7-14.

(8) Tuan, P. L.; Proctor, H. S.; and Couluris, G. J. Advanced Productivity Analysis Methods for Air Traffic Control Operations, Stanford Research Institute, Report DOT-TSC-FAA-76-27, December 1976.

APPENDIX D

SYSTEM PERFORMANCE CONSTANTS, DATA ELEMENTS, AND MEASURES

TRAFFIC SAMPLE CONSTANTS.

The constants describing a given traffic sample, with explanation, are as follows:

C1, NUMBER OF AIRCRAFT IN THE SAMPLE. This constant is derived automatically, by an unmanned running of the sample with the data reduction and analysis program operative. The sum of all aircraft programmed per time period is collected.

C2, IDEAL AIRCRAFT TIME IN SYSTEM. This is the total time that all sample aircraft exist in the simulation run period. This ideal time is totaled in an unmanned run; therefore no delays, speed reductions or controller actions are able to affect it. Since arriving aircraft are not descended, and departing aircraft are not climbed, this figures does not completely represent realistic ideal flying time, but is slightly shorter than that and is a retrievable amount, of a constant value, which relates to other samples in comparative terms.

C3, RATIO OF THE IDEAL AIRCRAFT TIME IN SYSTEM AND THE NUMBER OF AIRCRAFT IN THE SAMPLE. This consideration of the two constants described above (C2/C1) results in an average time during which each programmed aircraft exists in the simulation period. It is felt that this figure, since it considers over fast arrivals (not descended, therefore fast) balanced against over slow departures (not climbed, therefore slow), in an equally balanced sample, closely approximates a correct value.

C4, NUMBER OF COMPLETABLE FLIGHTS. This constant is derived in an unmanned run and is the total of programmed aircraft which have crossed the sector from incoming handoff point to outgoing handoff point, or from airport to outgoing handoff point, or from incoming handoff point to airport. All programmed flights which have not reached these described "completion" points at the end of the data period are not counted in this total figure. This total is also affected by the sterility of derivation explained above, but a balanced sample causes it to approach a correct (real) value.

C5, DATA PERIOD DURATION. A segmented data program has been constructed for use with this system. It is made up of 30-minute portions of the 3-hour traffic sample and combinations of these portions in the following format: (numbers 1 through 6 represent successive 30-minute portions to a total of 180 minutes).

```
1
1 2
2
1 2 3
3
1 2 3 4
4
1 2 3 4 5
5
1 2 3 4 5 6
6
      4 5 6
```

The segmentation is to serve as an investigative method for the final structuring of the correct sample size and duration.

C6, NUMBER OF ARRIVALS. Since each aircraft which will finish its flight at an airport or at a lower altitude in an adjacent arrival sector satisfies the definition of "arrival," and has to be descended and/or given an approach clearance, there is a requirement that the aircraft be at, or near, a procedural altitude. Therefore, the "arrival" is marked and has an accompanying "end altitude" (explained below). These are easily counted, first in an unmanned run, in which the prime task is descending all arrivals, without regard for proper control technique; thus an appropriate altitude is determined for each flight to finish.

C7, NUMBER OF DEPARTURES. Aircraft which will depart from an airport within the system, or, in certain en route samples, will have departed immediately adjacent to the problem sector, require climbing to optimum altitudes. A requirement exists that these altitudes be attained; therefore, the "departure" is marked and can be counted in unmanned running and in special manned runs wherein all departures are climbed without regard to proper control technique. These end altitudes are thence regarded as ideals to be attained.

C8, ARRIVAL/DEPARTURE RATIO. This ratio of the preceding two constants (C6/C7) can serve as one descriptor of the traffic sample. In later analyses, changing scores can probably be easily related to this indicator of traffic sample structure.

C9, ARRIVAL RATE SCHEDULED PER HOUR, and C10, DEPARTURE RATE SCHEDULED PER HOUR. These rates, easily determined from the appropriate data explained above, are sample descriptors and are especially meaningful when compared to actual rates achieved, which will be explained later.

BASIC PERFORMANCE DATA ELEMENTS.

The following describe the simple or first-order measures as taken or detected by the computer.

DE1, TARGET SPACING ANALYSIS--TERMINAL (3 NAUTICAL MILES (NMI)). For the phrase, "target spacing analysis," basically read here: "terminal area confliction." This setting of the parameter is appropriate for terminal ATC facility spacing; broadband radar within 40 nmi of the radar site. See also ATP 7110.65, paragraph 740. A score accrual occurs here when two targets approach within 3 nmi of each other at the same level (less than 1,000 feet separation). Also see DE5.

DE2, TARGET SPACING ANALYSIS--EN ROUTE (5 NMI). This is pairing of aircraft which are within 5 nmi of each other with less than minimum vertical separation (1,000 feet). This dimension is appropriate for en route ATC facilities. Radar site adaptation in today's narrow-band system is such that 5 nmi is practically universal, and the concept of discontinuing this spacing after targets have passed is no longer applied as it may continue to be applied in a broadband radar environment, i.e., ARTS III.

DE3, TARGET SPACING ANALYSIS--EXPERIMENTAL (10 NMI). Though this is the spacing applicable only for flights with less than 2,000 feet separation above flight level 600, the thinking for including it here was suggested by other similar work in traffic control analysis, known as the "index of orderliness." Ten nmi separation with less than minimum vertical separation is certainly an indicator of increasing workload and might relate to other measures in ways not known at this time. As an aside, this whole system of data collection has always been considered by its designers as not finished and subject to evolutionary changes as knowledge is gained.

DE4, TARGET SPACING ANALYSIS (TOTAL TERMINAL, EN ROUTE, AND EXPERIMENTAL) . This is summation of the three elements explained above (DE1+DE2+DE3) and reduced by those passing or diverging flights in terminal areas meeting special criteria (DE5).

DE5, TARGET SPACING ANALYSIS--TERMINAL (PASSING, AT LEAST 15°). This is a summation of apparent conflicts which satisfy the requirements of ATP 7110.65, paragraph 741:

"741. PASSING OR DIVERGING

Except in Stage A, vertical separation between aircraft may be discontinued when the following conditions are met:

a. You observe that they have passed each other or that one has crossed the projected course of another.

b. Their tracks are monitored to assure that their primary targets or beacon control slashes will not touch.

c. Their courses diverge at least 15°."

In the final treating of this data, therefore, those apparent conflicts (DE1) which can rightly be scored against a terminal (or other broadband user) are correctly diminished by this value.

DE6, DE7, and DE8. These were experimental data elements now deleted.

DE9, NUMBER OF DELAYS (START TIME). The controller subject of these tests is allowed to say to the external sectors, "I do not want any more traffic sent to me." A device for acceding to this request is to not allow new scheduled traffic to start. The number of these occurences is recorded.

DE10, DELAY TIME (START TIME). This is the duration of all the delays defined immediately above, expressed in minutes.

DE11, NUMBER OF DELAYS (HOLD AND TURN). This includes those delays to active targets caused by formal hold commands which result in "racetrack" flight patterns at a fix, and or any turn command which results in a turn of greater duration than 100 seconds. It was judgmentally decided that a turn of this magnitude would comprise more than a normal spacing turn or a turn along an airway. Therefore, all deliberate delaying turns, including simple "make-a-circle" type delays are detected. The value, 100 seconds, is subject to review and this value (approximately a 300° turn at lower altitudes) could be reevaluated after experience was gained.

DE12, DELAY TIME (HOLD AND TURN). This is the duration of all the delays explained immediately above, expressed in minutes.

DE13, NUMBER OF DELAYS (ARRIVAL). This is the number of above defined delays, start time and hold and turn, concerning arrival aircraft. All arrival aircraft are marked in the traffic sample list.

DE14, DELAY TIME (ARRIVAL). This is the duration of all the delays explained immediately above (DE13), expressed in minutes.

DE15, NUMBER OF DELAYS (DEPARTURE). This is the number of start time and hold and turn delays concerning aircraft marked as departures in the traffic sample.

DE16, DELAY TIME (DEPARTURE). This is the duration of all the delays explained immediately above (DE15), expressed in minutes.

DE17, NUMBER OF DELAYS (TOTAL). This is the sum of start time and hold and turn delays. It includes all delays accumulated during the test period.

DE18, DELAY TIME (TOTAL). This is the duration of all the delays, both start time and hold and turn, expressed in minutes.

DE19, AIRCRAFT TIME-IN-SYSTEM (REAL). This is the total duration of all aircraft on the "active" sector's frequency expressed in minutes.

DE20, NUMBER OF AIRCRAFT HANDLED. This is the total of all aircraft accepted by the active sector.

DE21, RATIO OF AIRCRAFT TIME-IN-SYSTEM (REAL) AND NUMBER OF AIRCRAFT HANDLED. This ratio of the two elements above, (DE19/DE20) amounts to the average time-in-system per aircraft and should be sensitive to performance by the subject controller over successive applications of the same test.

DE22, NUMBER OF COMPLETED FLIGHTS (TOTAL). This figure is the total of all aircraft which have been handled and which have either landed at an internal airport or completely flown the sector to a geographic point where they have been handed off to the next sector. This total differs from DE20, Number of Aircraft Handled, in that it is reduced by the number of aircraft which were left over in the sector at the end of the data period.

DE23, NUMBER OF ARRIVALS ACHIEVED. This is the total number of aircraft marked as arrivals which have completed their flights to the point of landing. It is especially meaningful in an approach control simulation, wherein the subject is asked to handle large numbers of arrivals.

DE24, ARRIVAL RATE ACHIEVED PER HOUR. This is the above figure (DE23) expressed as it relates to time. Here an increasing rate can directly indicate an improving performance, Target Analysis (DE1-DE5) considered.

DE25, NUMBER OF DEPARTURES ACHIEVED. This is the total number of aircraft marked as departures which have been enabled to takeoff.

DE26, DEPARTURE RATE ACHIEVED PER HOUR. This is the number of departures (DE25) per unit of time (hour). Here, an increasing rate can directly indicate an improving performance, Target Analysis (DE1-DE5) considered.

DE27, ARRIVAL ALTITUDES NOT ATTAINED. Each arrival in the sample is evaluated in that an ideal finish altitude is appropriate for each flight. In the case of an arrival at an airport, the airport elevation (ground level) is the proper finish altitude. In the case of an arrival meant for an airport in an adjacent sector, such that descent procedures are incumbent upon this subject's sector, the proper altitude is expressed as a range of altitudes, e.g., 9, 10, or 11 thousand feet, with the target altitude, of course, being the lowest one. Upon handoff or landing, each arrival is verified for proper range of altitude as specified in its traffic sample entry data. A count of flights not meeting this specification is made and this is considered a count of, in this sense at least, improperly handled aircraft, i.e., an "arrival not attained" count.

DE28, DEPARTURE ALTITUDES NOT ATTAINED. Each departure in the sample is evaluated, in that an ideal finish (requested) altitude is appropriate for each flight. Since several (2, 3, or 4) aircraft can possibly be proximate to each other, a range of altitudes is permitted to allow the subject controller to effect separation by using slightly different altitudes, but not largely different, from the requested altitude of the departure. A count of departed flights not reaching at, or near their requested altitudes, is compiled as a negative score, in that the departures have, in that sense, been improperly handled.

DE29, NUMBER OF AIR-GROUND CONTACTS (AK). This is a compilation of the number of times the subject controller has depressed his microphone switch. It is really a count of ground-to-air transmissions. It is undoubtedly a good indicator of controller organization and efficiency. "AK" is a telephone company acronyn meaning access key.

DE30, AIR-GROUND COMMUNICATIONS TIME. This is the total time in minutes that the controller has had his microphone switch depressed. This amount again is a measure of organization and efficiency when weighed against the work accomplished.

DE31, NUMBER OF ALTITUDE CHANGES. This is the number of times a climb or descent command was given during the data period. This would be a sensitive measure of workload and presumably would increase as traffic increased.

DE32, NUMBER OF HEADING CHANGES. This is the number of times that the aircraft in the sample were given "radar vectors" (i.e., the total count of verbal heading assignments made by the subject controller). This element has to be evaluated carefully (also applies generally to performance elements elsewhere) in that the general view that "many headings are a negative indicator" may not be entirely correct; many headings may be an unavoidable response to special procedural requirements within the simulation problem.

DE33, NUMBER OF SPEED CHANGES. This is the total number of times that the subject controller has verbalized a speed change to an aircraft in the sample. Generally, these are speed reductions. Again, a high count is interwoven with technique and necessity. The data must be weighed against all factors.

DE34, NUMBER OF PATH CHANGES (ALTITUDE, HEADING, AND SPEED). This is the total of all the elements described in the three places immediately above (DE31+DE32+DE33). Cautions relating to each also relate to this total.

DE35, NUMBER OF HANDOFFS. This is the total number of aircraft which have crossed the problem sector or departed from within the sector and have reached an exit point and are formally (through keyboard entry) transferred to another sector (in this simulation to a ghost sector or simulated sector). The lesser value of number of aircraft successfully handled and handed off would be less desirable, all other things considered.

DE36, HANDOFF DELAY TIME. This element is concerned with handoffs made by the ghost sector into the test sector. It is the summation of the time measured from the initiation of the handoff by the adjacent sector until its acceptance, by keyboard entry, by the subject controller. It is a sensitive measure of "busy-ness" and/or attentiveness.

DE37, RE-IDENTS. This concerns the command to a "pilot," within the system, to "ident," meaning activate the identification feature of the "airborne" transponder. The activation of this feature causes a reinforcement of the simulated target through flashing, which action simulates normal "identing" action sufficiently typical of the actual equipment in field use. Utilization of this feature in a "tagged" environment would signal target confusion caused by tag overlap and/or an overloaded controller reinforcing a deteriorating control situation.

COMPOUND PERFORMANCE MEASURES.

The following are compound measures derived from combining constants and data elements.

MI, HALF THE RATIO OF TERMINAL TARGET SPACING ANALYSIS TO AIRCRAFT HANDLED. The ratio of terminal conflicts (for "target spacing analysis" read conflict) per aircraft handled is corrected for pairs of aircraft in conflict by the division performed here (DE1-DE5/DE20 X .5). In other words, a two-aircraft conflict pair is the same as one conflict. Terminal spacing is 3 nmi.

M2, HALF THE RATIO OF EN ROUTE TARGET SPACING ANALYSIS TO AIRCRAFT HANDLED. The ratio of en route conflicts per aircraft handled is corrected for pairs of aircraft in conflict by the division performed here (DE2/DE20 X .5). Again the number is adjusted so that two aircraft represent one conflict. En route spacing is 5 nmi.

M3, HALF THE RATIO OF EXPERIMENTAL TARGET SPACING ANALYSIS TO AIRCRAFT HANDLED. This ratio (DE3/DE20 X .5) is also handled as explained immediately above and is thus reduced to similar terms so relative meaning can be retained. Experimental spacing is 10 nmi.

M4, TOTAL TARGET SPACING ANALYSIS OF ALL TYPES (TERMINAL, EN ROUTE, AND EXPERIMENTAL) PER THE DELAYS OF ALL TYPES, SCORED. This ratio (DE4/DE17) is a combination of two sensitive elements and is handled in this way for experimental purposes.

M5, DELAYS PER AIRCRAFT IN SAMPLE. All delays, both start time and hold and turn delays are averaged over the sample population (DE17/C1) so that the resulting value has a comparative meaning when weighed against another value from a different sample; i.e., one delay in a light sample is of more weight than one delay in a heavier sample, and this ratio tends to show this logic.

M6, RATIO OF DELAY TIME TO THE NUMBER OF AIRCRAFT IN THE SAMPLE PER HOUR. The total delay time (both start time and hold/turn) is apportioned over the aircraft population in the sample and adjusted so that it is expressed per unit of time (hour) (DE18/C1 per hour). Again this treatment of the measure tends to make the measure comparable over differing samples.

M7, COMPLETED FLIGHTS PER COMPLETABLE FLIGHTS. This is a comparison of the actual achievement of the subject controller; the number of flights which have completed their intended tracks versus the possible number of aircraft which were intended to be completed (DE22/C4). This value has a connotation of achievement, but it should be further evaluated in the light of the requirements of ATC. In other words, the population of the sample can achieve completeness without regard to required spacing of traffic; therefore, the value the subject attains may be less than blindly derived completable flights in an unmanned running of the sample, and a high level of accomplishment is nevertheless indicated.

M8, CONTACTS PER AIRCRAFT HANDLED. This measure is the average number of individual messages made by the subject controller to each aircraft he actually accepted into the simulation (DE29/DE20). It is probably a sensitive indicator of efficiency and also an indicator of mastery of the system. This latter idea has credence if we consider that if a controller knows what he is doing and has a firm idea of the problem ensuing, he will economically direct the aircraft on his frequency with a minimum of verbiage.

M9, COMMUNICATION TIME PER CONTACT. This averages the communication time over the number of messages made by the controller (DE30/DE29). Here again short messages are efficient.

M10, COMMUNICATION TIME PER HOUR. This measure (DE30 per hour) represents the amount of conversation a controller uses in a term which is sensitive to differing sample sizes and differing sector configurations. Caution is indicated in evaluating its meaning though some may think it transparent.

M11, RATIO OF AIRCRAFT HANDLED TO THE TOTAL NUMBER OF AIRCRAFT IN THE SAMPLE. This ratio (DE20/C1) is a measure of accomplishment. A calibrated sample would have an optimum potential which would immediately place a value upon an achieved score.

M12, CORRELATION, AND M13, TRANSFORMATION. These two measures represent the product-moment correlation coefficient computed on the basis of data points every 10 minutes using the total of all delay time and the total number of aircraft handled within a data hour. The r is transformed using the z function (DE20, DE17).

M14, RATIO OF AIRCRAFT TIME IN THE SYSTEM EXPERIENCED BY THE CONTROLLER AND THE AIRCRAFT TIME IN THE SYSTEM DERIVED IN AN UNMANNED RUNNING (REAL/IDEAL). This ratio (DE19/C2) reduces these figures to an easily compared average and is a measure of accomplishment.

M15, RATIO OF PATH CHANGES (THE TOTAL OF HEADING CHANGES, ALTITUDE CHANGES, AND SPEED CHANGES) AND THE TOTAL NUMBER OF AIRCRAFT HANDLED. This value is a measure of efficiency averaged over an amount of accomplishment (DE34/DE20) and should be a sensitive indicator of expertise which should differ markedly from individual to individual. Again caution is advised to prevent the reading of a wrong meaning into this mark. Air traffic control requirements in a certain sector having a certain aircraft population might dictate an optimum score here.

(It should be apparent to the reader that this system of measuring the performance of a controller is heavily dependent upon "calibrating" problems over a wide population of subjects.)

M16, RATIO OF ARRIVAL RATE ACHIEVED AND ARRIVAL RATE SCHEDULED. This ratio (DE24/C9) is especially applicable to terminal problems and as its value approaches unity, optimum performance is indicated. A calibrated sample is presupposed, and a value less than unity can still indicate optimum performance. Approach control expertise is the area of ATC investigated here.

M17, RATIO OF DEPARTURE RATE ACHIEVED AND DEPARTURE RATE SCHEDULED. This ratio (DE26/C10) is specifically aimed at a terminal controller's performance in the capacity of managing terminal departures. Remarks applicable to the arrival ratio mentioned above are also applicable here.

M18, AVERAGE INTERVAL OF ARRIVALS: M19, VARIANCE OF ARRIVAL INTERVALS: M20, AVERAGE INTERVAL OF DEPARTURES: AND M21, VARIANCE OF DEPARTURE INTERVALS. These four measures treat the core of a radar approach controller's functions. The emphasis on controller performance at any approach control facility is centrally concerned with these measures. The pressure of achieving efficient rates in these categories is ever present and tied to the facility's performance record. Indeed, here is the essential difference between en route traffic control and the terminal counterpart. It is also thought that the regularity which can be applied to this function is also germane to controller performance; hence, the separate measures examining the variance found in the series of control functions. It is felt that these measures would be prime achievement scores in a calibrated approach control problem.

M22, RATIO OF ARRIVAL DELAY TIME AND ARRIVAL DELAY EVENTS. This ratio (DE14/DE13) is a performance measure applicable to both en route and terminal tasks. Even though delays might be inevitable in a given sample, achievement is still detectable in an optimumly low delay average indicated here.

M23, RATIO OF DEPARTURE DELAY TIME AND THE NUMBER OF DEPARTURE DELAYS. This ratio (DE16/DE15) is also applicable to both en route and terminal performances. The logic used in describing the counterpart arrival measure is also appropriate.

M24, RATIO OF HANDOFF TIME AND THE NUMBER OF HANDOFFS. (DE36/DE35). A "handoff" is the transfer of control from one control position to another and the subsequent transfer of communication channels and radar tag information. The time taken to do this function is an indicator of both "busy-ness" and efficiency. It is felt that this value could be idealized in a calibrated sample.

TM-12-005
ATCSF-78-002

APPENDIX E

EXPERIMENTAL PLAN

FOR

CONTROLLER PERFORMANCE MEASUREMENT SCORING SYSTEM

CONTRACT NO. DOT-FA77NA-4011

Task Assignment No. 2

January 1978

Prepared by: T. E. Morgan, Jr., CSC

Prepared for the

NATIONAL AVIATION FACILITIES EXPERIMENTAL CENTER

ATLANTIC CITY, NEW JERSEY

COMPUTER SCIENCES CORPORATION
System Sciences Division
8728 Colesville Road
Silver Spring, Maryland 20910

ABSTRACT

This report provides the rationale, procedures, and analysis methods for an experiment designed to develop a quantitative method of measuring controller performance. The plan is based on the use of a standard work sample test. A selection of subjects will control aircraft in a fixed set of problems in a simulation environment. Each run will be evaluated by selected judges. The resultant data consisting of judges' scores and simulation measures will form the basis of an analysis using stepwise multiple regression techniques to estimate the judges' scores from the simulation measures.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

SECTION 1 - INTRODUCTION

The Controller Performance Measurement (CPM) activity is engaged in adapting and applying measures derived from NAFEC's Air Traffic Control Simulation Facility (ATCSF) to the evaluation of individual air traffic controller performance. Individuals, each of who handle the same traffic sample, will be compared and evaluated using the methods, techniques and measures developed in this activity.

Under Task 2 of Contract No. DOT-FA77NA-4011, Computer Sciences Corporation (CSC) has been assigned to support the CPM activity specifically in the areas of data analysis and development of computer program specifications. As a first step in CSC's support of CPM, this document was prepared to define the experimental procedures which will be used to collect data for subsequent analysis.

Previous work on the CPM activity has established the basic format and procedures for a work sample test based on digital simulation of the Air Traffic Control environment. In the immediate future, the experimental work will be devoted to the collection of a sufficient quantity of data to answer the following questions:

1) What is the best form of an automated grading system which will provide a single number index of controller performance as a result of testing?
2) At what traffic density level should the test be administered?
3) What is the minimum acceptable run length?
4) Is the test repeatable (i.e., does the same subject achieve a similar score on separate test sessions)?

5) Are the test results generalizable (i.e., does a subject's score on the test predict his performance on a range of other ATC situations)?

The following sections of this report describe a recommended plan for the collection of data to answer the above questions. Section 2 provides a background discussion on several issues pertinent to the data collection experiment. Section 3 describes the recommended approach and provides the rationale for design decisions. Section 4 describes the procedures to be employed in the analysis of data collected for this experiment, while Section 5 describes a series of contingency plans detailing subsequent work depending on possible experiment outcomes.

SECTION 2 - BACKGROUND

The plan for the next phase of CPM data collection is based on several major considerations including:

1. The facilities to be used.
2. The source and nature of the subjects.
3. The need for an external criterion against which to develop and validate a quantitative performance measure.
4. The expectation of a wide range of subject variations both in overall performance and in "style" of control.

2.1 TEST FACILITY

Subjects for CPM data collection will be tested using the digital Air Traffic Control Simulation Facility (ATCSF), previously the Digital Simulaation Facility (DSF), located at the National Aviation Facilities Experimental Center (NAFEC) in Atlantic City, New Jersey. The ATCSF provides a generalized Air Traffic Control simulation capability. A schematic of the facility is shown in Figure 2-1. An overview of the ATCSF's functional capability is provided in Digital Simulation Facility User's Guide (Reference 1). For CPM, the ATCSF will support simultaneous testing of four subjects. The additional controller positions are used for support (i.e., adjacent sectors) to the subject positions.

The ATCSF has been used in previous CPM experiments (see Reference 2 and 3) and software has been developed specifically for CPM to support independent

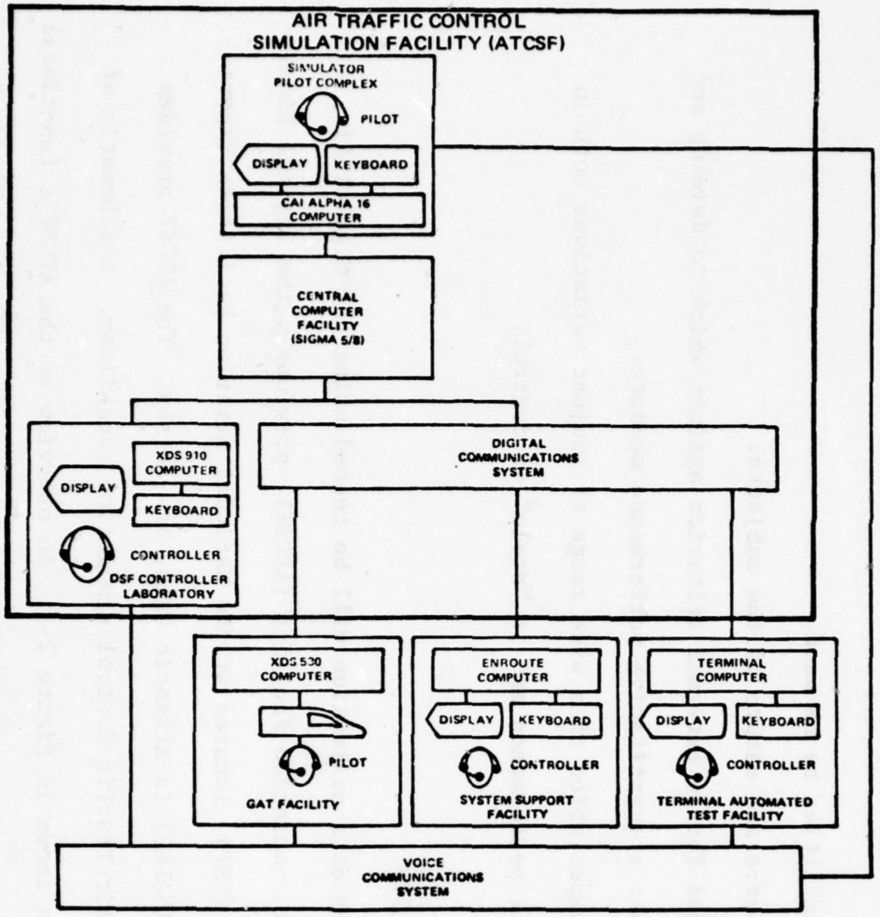


Figure 2-1. ATCSF Schematic

operation of the subject positions and to provide project oriented data reduction and analysis reports (see References 4). The only new software development effort required for this experiment is for a storage and retrieval system to provide efficient access to the large amounts of data collected (see Reference 5).

2.2 TEST SUBJECTS

Three options are available as sources of test subjects for the CPM data collection experiment. These options are: 1) NAFEC Evaluation Group, 2) Full Performance Level Controllers, and 3) Developmental Controllers.

The NAFEC Evaluation Group would provide a very convenient source of subjects. Unfortunately, the group is composed of too small and homogeneous a collection of individuals. The group consists of between 20-30 persons all with extensive ATC experience and several years of NAFEC simulation experience. As subjects, they are "simulation wise" and generalization of their performance for broader application would be highly questionable. (In typical NAFEC simulations, these subjects generally provide relative comparisons between system alternatives where absolute measurement of performance is not a requirement).

Use of either journeyman and/or developmental controllers from field facilities incurs a very large transportation and per diem expense along with disruption of work and training schedules at the facilities. Choice of subjects must take into consideration both the goals of the effort and the cost of bringing to NAFEC a sufficiently large number of subjects to provide a data base of statistical significance.

Development of a test to predict the future capabilities of a developmental controller requires that three actions be taken:

1. A test must be constructed which discriminates the performance of developmental controllers.
2. A test or method must be developed which provides a consistent overall evaluation of journeyman proficiency.
3. Longitudinal data must be collected on a sample of individuals from completion of academy training through their progression to certified journeyman controllers to relate performance as a developmental with subsequent performance as a journeyman.

Ideally, a study such as the one depicted in Figure 2-2 should be performed. The method shown would provide for development of a test for each experience level initially. The subjects within each experience level group would be retested in a series of longitudinal studies. In this manner, an initial approximation to the long term growth pattern could be determined after the first retest. Subsequent retests would further enhance the understanding of the controller growth process.

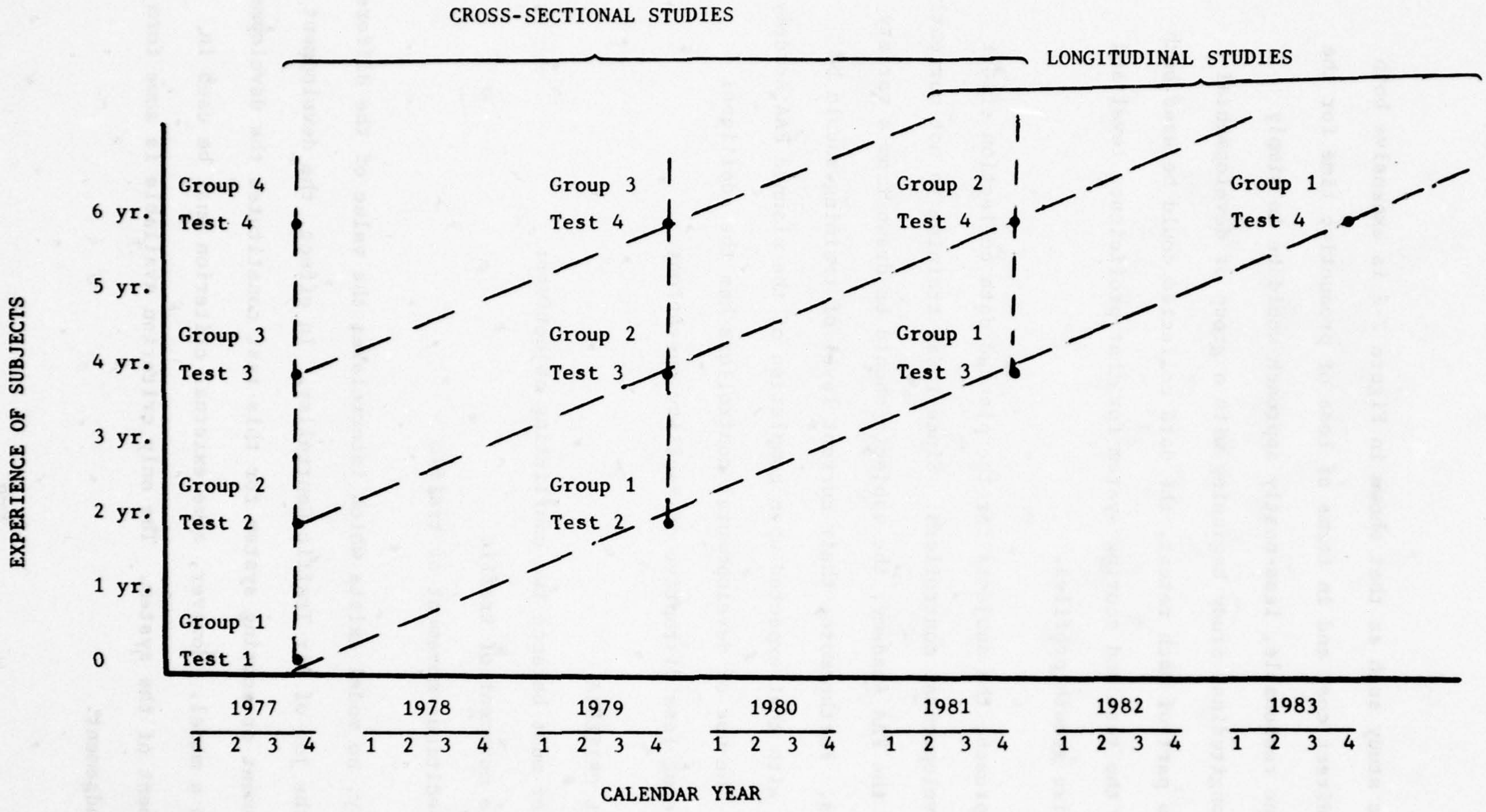


Figure 2-2. Ideal Methodology for Development and Calibration of Controller Performance Tests

An exhaustive study such as that shown in Figure 2-2 is expensive both in terms of direct cost and in terms of loss of productive time for the subjects. One reasonable, less-costly approach would be to simply initiate a longitudinal study beginning with a group of developmental subjects. As part of each retest, the data collected could be used both to calibrate the test and scoring system for that proficiency level and to characterize growth profiles.

With this approach, the subjects for the planned data collection effort should be developmental controllers. Since radar training is not currently conducted at the FAA Academy, the subjects should be drawn from a variety of facilities. Furthermore, their current level of training should be commensurate with that expected upon completion of the planned FAA Academy curriculum. The use of developmental controllers has the additional benefit of being less disruptive of facility operations.

2.3 EXTERNAL CRITERIA

The controller must balance two conflicting objectives:

1) Safe movement of traffic

2) Expeditious movement of traffic

Unfortunately, no model exists which interrelates the value of the different aspects of the job of Air Traffic Controller. In effect, the development of a measurement or scoring system for this test constitutes the development of just such a model. However, some external criterion must be used in the development of the system. The only criterion available is some form of expert judgement.

Two external judgements are readily available on each subject. These are 1) his grade at the FAA Academy and 2) his proficiency ratings by his training officer or supervisor. For the purposes of the planned CPM effort, FAA grades and supervisory ratings should be collected and used in analysis. However, it is recommended that a group of experts be selected to view simulation runs and judge the performance of subjects. These experts should be selected from field training officers and FAA Academy instructors with a substantial background in evaluating or certifying trainees. The judgements of these experts will be based on over-the-shoulder evaluation and will be collected in a controlled fashion so that individual biases can be quantified and considered in the modelling effort.

2.4 SUBJECT VARIATIONS

Numerous experiments have shown considerable performance differences between controllers. In particular, previous CPM studies (References 2 and 3) have shown large differences even within a fairly homogeneous sample such as the NAFEC Evaluation Group. It is anticipated that variations within a sample of developmental controllers will be even larger. For this experiment, the major concern is that a broad range of capabilities be sampled to provide insight into the entire trainee population. The only assurance of broad coverage is large sample size. In this case, the selection of subjects can be aided by using supervisory ratings which would provide some indication of performance. Subjects should be selected so as to insure a range of ratings from each facility.

From the standpoint of the experimental design, an even more important consideration is the variability of the expert evaluation. The development of a quantitative measure of performance depends on reasonable consistency within each individual and among the judges as well as the degree to which simulation measures reflect those judgements. This uncertainty affects the development of the experimental plan in two ways. First, the sample size must be fixed fairly conservatively, but with recognition of the resources available. Second, the plan must be devised so that additional samples can be collected without a major impact on the analysis.

SECTION 3 - RECOMMENDED APPROACH

The goals of this data collection effort are to determine:

1) What is the best form of a grading system which will provide a single index number as a result of testing?

2) At what density level should the test be administered?

3) What is the minimum acceptable run length?

4) Is the test repeatable (i.e., does the same subject achieve a similar score on separate test sessions)?

5) Are the test results generalizable (i.e., does a subject's score on the test predict his performance on a range of other ATC situations)?

3.1 IMPLICATIONS OF GOALS

To establish an experimental plan for the collection of data to satisfy the stated goals, a clear understanding of the implications of each question must be obtained.

3.1.1 Grading System

This goal can be restated as the development of a function of real-time simulation measures which reflects the rating of the judges. Previous work (Reference 6) has established tentative relationships must be investigated further and the efficacy of other simulation measures must be considered.

To this end, a large collection of simulation data and corresponding judgemental evaluations is necessary to perform this analysis. The only meaningful constraint on data collection for this purpose is that a variety of different combinations of judges be used to provide a comparison of different judges' evaluation criteria and standards.

3.1.2 Density Levels

Figure 3-1 provides some insight into the factors affecting density level selection. The basic premise of the figure is that there is some low density level at which all subjects will perform well and, similarly, some high density level at which all subjects will perform poorly. Since at those levels the majority of subjects would do either very well or very poorly, it would be difficult or impossible to discriminate between individuals. Between those levels, a broader range of performance is expected. In effect, this objective can be considered a search for the density which provides the maximum subject variation. Therefore, this experiment must collect data over a range of different traffic density level conditions.

In a previous experiment (Reference 2) conducted using NAFEC Evaluation Group controllers, three density levels (40, 50, and 60 aircraft/hour) were employed. Although that experiment was conducted using a small group of six subjects, the results indicate that the choice of density level is not too sensitive. However, since this experiment will use trainees instead of journeymen, it is anticipated that they will not be able to handle the same density of traffic. Therefore, the location of the peak variation (between 50 and 60 aircraft/hour) is only useful as an upper limit.

One other consideration in the investigation of density levels is the nature of the analyses which will be conducted. Hopefully, the density can be selected from one of those used in the experiment. However, if none of

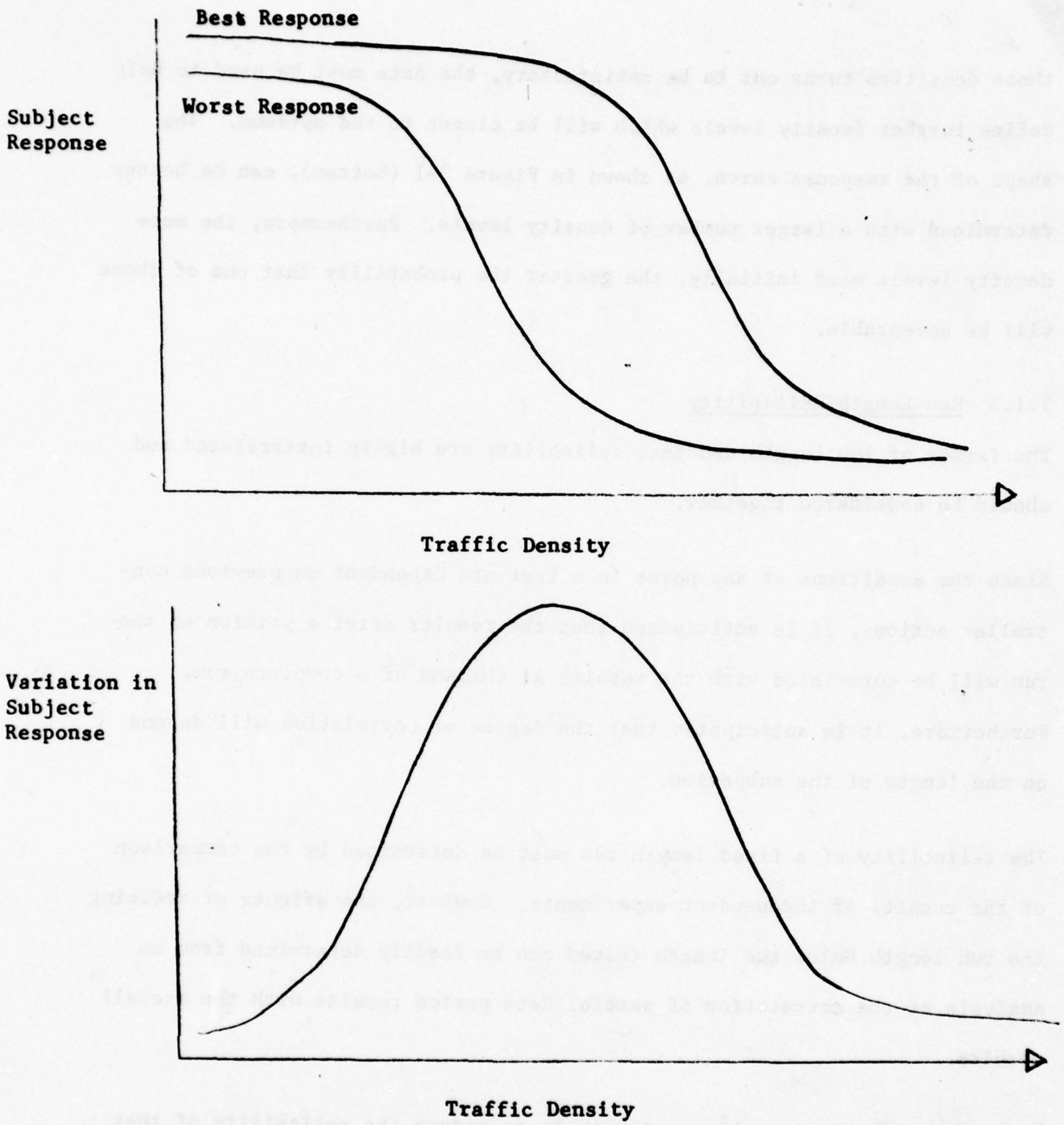


Figure 3-1.

Expected Subject Response to Varying Density Levels those densities turns out to be satisfactory, the data must be used to help define further density levels which will be closer to the optimum. The shape of the response curve, as shown in Figure 3-1 (bottom), can be better determined with a larger number of density levels. Furthermore, the more density levels used initially, the greater the probability that one of those will be acceptable.

3.1.3 Run Length/Reliability

The issues of run length and test reliability are highly interrelated and should be considered together.

Since the conditions at any point in a test are dependent on previous controller actions, it is anticipated that the results after a portion of the run will be correlated with the results at the end of a complete run. Furthermore, it is anticipated that the degree of correlation will depend on the length of the subperiod.

The reliability of a fixed length run must be determined by the comparison of the results of independent experiments. However, the effects of reducing the run length below the length tested can be readily determined from an analysis of the correlation of partial data period results with the overall results.

The effect of reducing the run length is to reduce the reliability of that single measurement. This reduction can be offset by making several independent runs of a fixed shorter test period. In fact, since one is dealing with multiple independent estimates, this is most likely a more effective score estimation approach. The question of test reliability can then focus on how many such independent sessions are needed to obtain the desired reliability.

For the purposes of this experiment, a run length of one hour has been selected. It is felt that one hour is the maximum permissable before fatigue or boredom set in. It is also felt that the optimum tradeoff point between run length and number of repetitions will occur with test lengths of less than one hour.

3.1.4 Test Generality

Certain skills are obviously necessary for a person to be a good air traffic controller. Individual differences do exist and, in fact, the motivation behind this experiment is to measure in some sense the degree to which particular individuals possess those skills. It is less clear, however, whether or not the same skills are required to the same degree in every air traffic control situation.

A work sample test such as a simulation test must be administered using a preestablished, fixed sector geometry and traffic load. The general applicability of the results of that test to the spectrum of possible ATC situations must be investigated and hopefully demonstrated. To this end, subjects must be tested in alternative sector geometries to determine whether or not the results of one test predict, in some sense, the results of the others.

3.2 LOGISTICAL CONSIDERATIONS

Logistical considerations impose limits on the order and number of runs which can be conducted. Travel and per diem expenses dictate that each subject be brought to NAFEC once and that maximum use be made of the subject during his stay. Considering the impact of training schedules and the intense nature of the testing, it is felt that each subject should spend at most one week as a subject at NAFEC.

Based on the current ATCSF schedules, it is unlikely that more than 1/2 day testing will be possible. In four hours, each subject could make a maximum of 3 one hour runs for a total of 15 runs in his one week stay.

3.3 EXPERIMENTAL DESIGN

The goals of this experiment as discussed in Section 3.1 lead to consideration of a factorial design with the following factors:

1) Traffic sample density levels

2) Test geometries

3) Student trials

Based on the logistical considerations discussed in 3.2 each subject will be tested in no more than 15 runs. This constraint limits the number of levels at which each subject can be tested.

With these considerations, the recommended structure of the experiment consists of three density levels, two test geometries, and two trials per subject for a total of 12 runs. These runs could be accomplished in four half days. Using only 12 runs provides a buffer against DSF system failure and other contingencies.

It is recommended that three density levels be used. The choice of three levels represents a reasonable compromise between increasing the size of the current test and the risk of having to perform further experimentation to find an appropriate density level. It is further recommended that the density levels be set at 35, 40, and 45 aircraft per hour. These settings cover a range from the lower density at which journeymen controllers exhibited reasonable variation to about half the maximum level at which their variation was acceptable.

A basic test of generalizability can be conducted using only two geometries as long as the differences between those geometries is significant in an operational sense. While using two levels will not demonstrate complete generality, it will indicate whether or not meaningful differences in controller skills are required in different tactical situations. It is recommended that the two geometries employed in the experiment be representative of a high-altitude and a transition sector.

Similarly, two replications or applications of each test will be adequate to obtain a reasonable estimate of test reliability. In effect, each subject will provide six measurements of reliability, one for each different density and geometry level. The relationship between test reliability and the other factors can, therefore, be studied.

3.4 OTHER DESIGN CONSIDERATIONS

A number of other considerations effect the details of planning the experiment. These considerations include the number of judges, the sequence in which the tests will be conducted, and the number of subjects to be used.

The number of judges to be used is limited by the number which can effectively watch a single test session without being overly obtrusive. At least two judges per subject is required to permit the cross-comparison of judges and, practically, that is the most judges which could observe each subject. Therefore, a total of eight judges is recommended. These judges should remain the same throughout the experiment.

The use of eight judges would provide 28 distinct pairs or combinations of judges to evaluate any run. To provide a cross comparison of judges' evaluation criteria and standards, a reasonable selection of the pairs should be employed.

The sequence with which a subject experiences his runs may well affect his performance on those runs. Therefore, due care must be taken to insure that the effects of run sequence are distributed equitably along the factor levels.

One reasonable approach to organizing the test sequence is to test each subject on all three density levels within a single day. The order in which the subject experiences each density level within the day can be controlled and the order in which the subject experiences the geometries and replications over the four days of tests can also be controlled.

Table 3-1 gives a complete enumeration of the possible density and geometry/replication sequences. In each case, a total of six sequences are possible. Each subject will experience four of the six density sequences and one of the geometry/replication sequences. Six subjects are required to balance the geometry/replication sequence and 15 are required to balance the density sequences.

Groups of 30 subjects would, therefore, be required to balance all sequence effects. Since there are no currently available data describing the variation expected, it is impossible at this time to specify the exact number of subjects required. The experiment would be conducted in a sequential fashion stopping after each set of 30 subjects to determine if additional data is required. It is expected that 2-3 blocks of 30 subjects will prove adequate.

3.5 COLLECTION PLAN

Control of data collection depends on establishing the proper sequences for each subject and each judge. Figure 3-2 gives the positions the judges should occupy for each run of the data collection effort. Subject number N receives geometry replication sequence G_j(see Table 3-1) where j = N modulo 6 (hereafter denoted as N mod 6) and density combination C_k (see Table 3-2) where k = N mod 15. The actual density sequences are randomly selected from all combinations of the particular combination used.

Density Sequences:

$d_1\ d_2\ d_3 = D_1$

$d_1\ d_3\ d_2 = D_2$

$d_2\ d_1\ d_3 = D_3$

$d_2\ d_3\ d_1 = D_4$

$d_3\ d_1\ d_2 = D_5$

$d_3\ d_2\ d_1 = D_6$

Geometry/Replication Sequences:

$g_1\ g_1\ g_2\ g_2 = G_1$

$g_2\ g_2\ g_1\ g_1 = G_2$

$g_1\ g_2\ g_1\ g_2 = G_3$

$g_2\ g_1\ g_2\ g_1 = G_4$

$g_1\ g_2\ g_2\ g_1 = G_5$

$g_2\ g_1\ g_1\ g_2 = G_6$

d_1 = 35 aircraft/hour

d_2 = 40 aircraft/hour

d_3 = 45 aircraft/hour

g_1 = high altitude sector

g_2 = transition sector

Table 3-1

Density and Geometry/Replication Sequences

	Judges Occupying Position			
Run Index Number [1]	Simulator Position 1	2	3	J
1	J_1 J_2[2]	J_3 J_4	J_5 J_6	J_7 J_8
2	J_1 J_3	J_2 J_4	J_5 J_7	J_6 J_8
3	J_1 J_4	J_2 J_3	J_5 J_8	J_6 J_7
4	J_1 J_5	J_2 J_6	J_3 J_8	J_4 J_7
5	J_1 J_6	J_2 J_5	J_3 J_7	J_4 J_8
6	J_1 J_7	J_2 J_8	J_3 J_6	J_4 J_5
7	J_1 J_8	J_2 J_7	J_3 J_5	J_4 J_6

[1] Index number = data run number mod(7)

[2] J_i indicates judge number i, i.e., judge 1, etc.

Figure 3-2. Judge Position Control

C_1	=	D_1 D_2 D_3 D_4[1]
C_2	=	D_1 D_2 D_3 D_5
C_3	=	D_1 D_2 D_3 D_6
C_4	=	D_1 D_2 D_4 D_5
C_5	=	D_1 D_2 D_4 D_6
C_6	=	D_1 D_2 D_5 D_6
C_7	=	D_1 D_3 D_4 D_5
C_8	=	D_1 D_3 D_4 D_6
C_9	=	D_1 D_3 D_5 D_6
C_{10}	=	D_1 D_4 D_5 D_6
C_{11}	=	D_2 D_3 D_4 D_5
C_{12}	=	D_2 D_3 D_4 D_6
C_{13}	=	D_2 D_3 D_5 D_6
C_{14}	=	D_2 D_4 D_5 D_6
C_{15}	=	D_3 D_4 D_5 D_6

1) The coding for D_i is shown in Table 3-1.

Table 3-2

Density Combinations

SECTION 4 - ANALYTICAL APPROACH

Analysis of data from this experiment will proceed in a series of steps:

1) Development of scoring system
2) Determination of reliability/run length
3) Selection of density level
4) Validation of scores
5) Evaluation of generalizability

Each of these analyses is discussed in the following sections.

4.1 DEVELOPMENT OF SCORING SYSTEM

As data are collected, an investigation of the relationship between judges' evaluations and a variety of simulation measures will be conducted. In this first effort, data from only one of the replicates for each test condition will be used. The data from the second replicate will be saved for validity tests (see Section 4.3).

This first effort can be characterized as an exploratory data analysis in the sense that the form of the relationship between judgements and the simulation measures is largely unknown. Previous work (Reference 2, 3, and 6) developed and used certain preliminary forms of this relationship. These relationships will be used as a point of departure and several new simulation measures proposed (Reference 7) will be considered in estimating judges scores.

Plots of the simulation measures versus judges' evaluations will be generated from the data with each density level and geometry. The general form of these separate relationships will be used to select a small subset of measures (< 15) which can form the basis for a model construction.

Stepwise multiple regression techniques will be employed to find the best estimator of judges' scores from the selected simulation measures. This technique (see Reference 8, page 171) provides a means of efficiently examining alternative model forms composed of an increasing number of independent variables. The result of this procedure will be a model of the form

$$Y = \beta_0 + \beta_1 f_1 (X) + \beta_2 f_2 (X) + . . . + \beta_n f_n (X)$$

where

Y = the set of judges scores

$\beta_0, \beta_1, \beta_2, . . . \beta_n$ = estimated regression coefficients

X = matrix of simulation measures collected

$f_i(X)$ = transformation functions of the collected measures

Variations of the selected measures will be examined and, if necessary, new simulation measures will be defined in an attempt to discover an estimating relationship which is maximally correlated with the judges' scores while at the same time provides a justifiable working description of the components of controller performance.

In the sense of Cohen (Reference 9, page 253), this analysis is to be conducted without the luxury of a prior theoretical model upon which to base judgements of model validity. Therefore, due care must be exercised to insure that models developed using multiple regression techniques are operationally meaningful.

The highly correlated nature of many of the simulation measures causes certain problems which must be guarded against. As discussed in Reference 9, this problem known as multicollinearity may cause problems in the interpretation of results and in the stability of regression coefficients. The best protection available for these problems is to prescreen the transformed variables and eliminate all but one of each set of highly correlated variables. This effort will be conducted independently for each density level and geometry and, in addition, a single estimator will be developed for each density without regard to geometry.

4.2 RELIABILITY/RUN LENGTH

For each density level, the scores from the start of the problem up through each five minute subperiod will be computed and the functional relationship between subperiod length and the intraclass correlation (Reference 10, page 16) between the scores for the two replicates will be plotted. Using this information, the number of test runs of differing lengths required to yield an overall test reliability of .9 will be determined. The run length selected will be that subperiod requiring the minimum total test time.

4.3 SELECTION OF DENSITY LEVEL

Results of the preceeding analyses will be used to select the density level. The correlation between judges' scores and the estimating relationship values will be plotted as a function of traffic density to obtain an overall impression of the relationship. Similarily, the number of runs required to achieve a .9 reliability will be plotted against density level. Selection of the density level will depend on both the reliability and quality of the estimating relationships developed.

Assuming acceptable correlation and reliability is achieved in at least one level, then either of two courses of action will be taken. If a single density level produces a significantly better relationship, that level will be chosen. If the response is reasonably equal or flat, the middle density level will be used in subsequent tests. However, if none of the relationships is acceptable, the shape of the density/correlation function will be used to direct further experimentation (see Section 5 - Contingency Plans).

4.4 VALIDATION OF SCORES

Once a scoring system and density level have been chosen, the remaining, unused data will be used to validate the scoring function. Each reserved run will be evaluated using the scoring function and the validity of the function will be tested by:

1) Determining if the correlation between the judges' evaluations and the scoring function is at least equal to the correlation within judges' scores.
2) Determining if the order of the subjects as ranked by the scoring system is the same as the rank order as determined from the judges' evaluations.

This analysis will be performed separately for each geometry.

4.5 EVALUATION OF GENERALIZABILITY

After the scoring system has been validated within each geometry, the relationship between the geometries will be tested to determine the effects of geometry. This analysis will proceed in a manner similar to the validation tests. Both the correlation between judges' evaluations and the scoring function and the constancy of subject rank will be tested to determine if these factors remain constant across geometries.

SECTION 5 - CONTINGENCY PLANS

The basic structure of this data collection effort was planned on the assumption that certain crucial conditions will exist. For example, it is assumed that an adequate density level selection can be made from the three density levels used in the experiment. It was further assumed that the scoring system developed will demonstrate reasonable reliability and generality characteristics. While previous work (Reference 2, 3, and 6) indicate that these conditions will occur, provision must be made for modifying the data collection plan in the event one or more of these conditions fails to occur. This section discusses some of the possible contingencies which might arise to invalidate these assumptions along with the impact of those contingencies on the data collection plan.

5.1 DENSITY LEVEL SELECTION

As data are collected, the variation between subjects will be monitored. The variance between subjects (as indicated by judges' scores) is indicative of the maximum discrimination between individual subjects. For example, if all subjects achieve the same ratings by the judges, it will be impossible to determine a function of real time simulation measures which accurately estimates the underlying scoring mechanism. Plots of this variance as a function of run density will be maintained and used to determine if the density levels being used are appropriate.

If it appears that the optimum density level is not included in the levels being used, the density levels will be changed appropriately and the experiment will, in effect, be restarted. The organization of the data collection effort, where each subject completes all his runs in a single week will permit a good estimate of the acceptability of the density levels used after approximately one quarter to one half of the subjects have been tested.

5.2 TEST RELIABILITY

Perhaps the crucial consideration in the development of a work sample test of controller performance is the question of test reliability. Test reliability is in fact a measure of variation of an individual in his performance. Large run to run variations in an individual's performance is indicative of serious system problems. However, with developmental controllers, larger run to run variation may reasonably be expected.

Although it is anticipated that the work sample test should prove to be sufficiently reliable, large run to run variations which resulted in a requirement for a large number of replications to obtain reasonably high reliability would be devestating for test administration. If this should occur, a detailed analysis of the failure modes of individuals must be conducted to determine if it is possible to identify aberrant runs and their causes. If the aberations are detectable, a scheme for requiring reruns can be developed and, perhaps, the overall test reliability can be increased. Otherwise, either the high cost of testing must be accepted or the concept of work sample testing must be rejected.

5.3 TEST GENERALITY

If controller performance as indicated by judges evaluations should be dependent on the particular scenario employed, further research will be required. In effect, this result would indicate that certain individuals are better candidates for one type of assignment than they are for some other.

If possible, predictive relationships will be developed which relate modes of performance on one test with performance on the other test. If these relationships can be successfully developed, than a single test can be administered and the individual's preferred assignment deduced. Otherwise, a battery of tests must be developed in order to identify individual tendencies. In either case, more data must be collected to examine controller performance over a broader spectrum of tactical situations.

REFERENCES

1. Digital Simulation Facility User's Guide, DOT/FAA, Simulation and Analysis Division, System Development Branch, June 1975.

2. Buckley, E. P., et al., Development of a Performance Criterion for En Route Air Traffic Control Personnel Research Through Air Traffic Control Simulation: Experiment I - Parallel Form Development, FAA-RD-75-186, DOT/FAA/NAFEC, 2/76.

3. Buckley, E. P., et al., CPM PROBE Experiment On Performance Information Feedback, NA-77-18-LR, DOT/FAA/NAFEC, Atlantic City, 4/77.

4. Algeo, R. and Pitale, J., Program Design Specifications for CPM Experimentation (Terminal and En Route), ATCSF-77-015, DOT/FAA/NAFEC.

5. Morgan, T., Requirement Specification for Data Storage and Retrieval System, TM-12-002, Computer Sciences Corporation, 8/77.

6. Buckley, E. P., et al., A Comparative Analysis of Individual and System Performance Indices for the Air Traffic Control System, NA-69-40, DOT/FAA/NAFEC, Atlantic City, 1969.

7. Young, E., Automated Measurements of Air Traffic Controller Performance in a Simulated Environment, Computer Sciences Corporation, TM-12-004, December 1977.

8. Draper, N. R. and Smith, H., Applied Regression Analysis, John Wiley and Sons, Inc., New York, 1966.

9. Cohen, J. and Cohen, P., Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences, Lawrence Erlbaum Assoc., 1975.

10. Des Raj, Sampling Theory, McGraw Hill Book Co., New York, New York, 1968.

APPENDIX F

CONTROLLER PERFORMANCE MEASUREMENTS IN SYSTEMS TEST AND EVALUATION

AD-A058 082   NATIONAL AVIATION FACILITIES EXPERIMENTAL CENTER ATL--ETC F/G 5/9
DEVELOPMENT OF A PERFORMANCE CRITERION FOR AIR TRAFFIC CONTROL --ETC(U)
JUL 78 E P BUCKLEY, K HOUSE, R ROOD

UNCLASSIFIED   FAA-NA-78-9   FAA-RD-78-71   NL

2 OF 2
ADA
058082

END
DATE
FILMED
10-78
DDC

COMPUTER SCIENCES CORPORATION

SYSTEM SCIENCES DIVISION (609) 641-8200
P.O. BOX 737 POMONA, NEW JERSEY 08240

March 1, 1978

Dr. Edward P. Buckley, ANA-230
NAFEC
Atlantic City, New Jersey 08405

Subject: Implications of CPM Data Collection on System Test and Evaluation Experiments

Dear Ed:

Based on my experience in working as a consultant on the experimental design of simulation experiments for many system test and evaluation efforts, I believe that the evaluation criterion to be developed and much of the data to be collected in your proposed experimentation for the Controller Performance Measurement (CPM) project would be of great value in the system test and evaluation experiments conducted at NAFEC. Furthermore, with certain extensions to the CPM effort, the planning and analysis of the system test and evaluation efforts could be placed on a sound scientific basis. Although my personal experience has been largely limited to the experiments conducted using the Air Traffic Control Simulation Facility (ATCSF), I'm sure that similar benefits would be accrued for experimentation on the Terminal Area Test Facility (TATF) and the System Support Facility (SSF).

Among the difficulties plaguing system test and evaluation efforts, two major problems stand out: A lack of planning data on which to base an experimental design and an excess of often contradictory performance measures. Data and results from the proposed CPM experiment would help alleviate both problems.

Planning an Experiment

As currently conducted, planning for a system test and evaluation experiment is more of an art than a science. Experimental designs are frequently dictated solely by available resources. The only role design plays is to insure that the data is collected in an unbiased fashion amenable to analysis by available statistical techniques. The major benefit that experimental design should provide -- the quantification of the relationship between the cost of experimentation and the risks of obtaining misleading results -- is ignored due to lack of data describing the error or variation expected in the experiment.

Most ATC experiments conducted at NAFEC are planned to determine the impact of proposed changes or additions on the system. To this end, these experiments normally culminate with statements concerning the presence or absence of significant changes in the system performance measures. As you are well aware, in any statistical analysis two types of errors can be made in statements of this type. The analysis may indicate that there was an effect when, in fact, the change or addition would not cause an impact or the analysis may indicate that there is no impact when there would be one. Both of these error conditions have implicit costs in the resultant decisions which should be investigated during the planning stage and balanced against each other and the cost of additional data collection.

The first error type, that of stating there is an impact when none exists, is generally called the type I or α error. This error can be arbitrarily established and for NAFEC ATC experiments it has generally been fixed at .05 on the basis of precedent.

The second error type, that of failing to detect the presence of an impact, is referred to as the type II or β error and is much more complicated to establish. Quantification of the type II error requires a knowledge of variation introduced by the different elements of the experiment such as controller subjects, traffic samples or sector types. Furthermore, the type II error is not a constant but depends on the magnitude of the actual impact and the choice of the type I error. Due to the lack of information concerning the expected variation, consideration of the type II error is typically ignored.

Lack of consideration of the type II error leads to experiments which are under-designed. In the NAFEC environment this unfortunate consequence arises from attempts to get as much as possible for the time and money spent. There is a tendency to include consideration of additional factors in each experiment. Statements like, "How does the equippage level (or traffic density or ...) affect the system response? Maybe we ought to throw in a couple of different equippage levels?", arise frequently. Without the ability to make quantitative statements about the effects of incorporating these factors, it is normally very difficult to stave off attempts to consider these very important but not necessarily central questions. The effect of incorporating these factors, however, is an overall increase in the type II error and consequent diminution of the chances of finding significant effects in any of the factors.

The need for planning data describing the expected experimental variation has been recognized historically. One attempt was made several years ago to establish a statistical data base which would store the results of ATCSF experiments and subsequently provide a source of data to estimate the expected experimental variation. This attempt was thwarted by the characteristics of the ATCSF experiments themselves. Each experiment is typically unique, with its own geographic scenario, traffic samples and subject organization. In addition, each experiment includes runs in which the system is modified by the incorporation of the change being studied. These differences create major difficulties in pooling the data from a group of experiments in order to obtain the needed information. A much higher degree of standardization of the data collected would be required before the expected experimental variation could be estimated by that technique.

The proposed CPM experiment would provide an excellent source of planning information. Since a large quantity of data would be collected in a consistent, standardized fashion, accurate estimates of the variation introduced by controller subjects could be obtained for use in subsequent planning. The proposal would provide for a picture of both variations within a subject and group variation. Expansion of the basic proposal to provide for a more complete investigation of the traffic density (load) and sector design characteristics would yield a very thorough picture of the error environment for the experimental factors most frequently used.

Analysis of Results

The ATC system is saddled with dual objectives: The safe and expeditious movement of aircraft. The achievement of each of these objectives is usually measured by a number of proxy variables such as confliction counts and average separations or aircraft time in system and delay. Many system test and evaluation efforts yield ambiguous results in which many of the measures employed are counterindicative, leading to one conclusion for one objective and to the opposite conclusion for the other objective. The most important aspect of system performance, the balance between safe and expeditious traffic movement, i.e., "Good Air Traffic Control", is buried amidst the detail of the many other indicators.

The absence of an overall indicator of system performance places the burden of resolving any conflicting indications on each individual experimenter. Since no standard or model exists for tradeoffs between safety and expeditious movement, the judgement of the experimenter must be empoyed to reach overall system conclusions. The conclusions reached in different evaluations thus include the personnal biases and prejudices of the individuals performing the analysis and the conclusions will not reflect any consistent rationale. Furthermore, the judgement used in any analysis may be deeply imbedded in the analysis employed making it difficult for a decision maker to assess the quality of the experimenter's judgement in this vital area.

A consistent, overall system performance measure would significantly improve the current approach to the analysis of ATC experimental results. The performance measuring system to be developed in the CPM experiment could fill that need. Although CPM is concerned with the evaluation of individuals performing the job of Air Traffic Controller within the confines of an existing system framework while system test and evaluation is concerned with assessing the impact of modifications and additions to the ATC system, both processes entail the measurement of how well the combination of controller and system affect aircraft movement.

Since the proposed CPM effort is aimed toward developing a performance measure based on the resultant aircraft motion, that scheme should be equally applicable to system test and evaluation experiments.

Sincerely yours,

COMPUTER SCIENCES CORPORATION

Thomas E Morgan

Thomas E. Morgan, Jr.
Project Director

TEM:djs

cc: C. Falkner
K. House
S. Pszczolkowski